

Internet traffic characterisation: Third-order statistics & higher-order spectra for precise traffic modelling

A.K. Marnerides^{a,*}, D.P. Pezaros^b, D. Hutchison^a

^aInfoLab21, School of Computing & Communications, Lancaster University, UK

^bSchool of Computing Science, University of Glasgow, UK



ARTICLE INFO

Article history:

Received 3 March 2017

Revised 27 January 2018

Accepted 31 January 2018

Available online 7 February 2018

Keywords:

Internet traffic characterisation

Traffic engineering

Higher order spectra

ABSTRACT

Undoubtedly, the characterisation of network traffic flows is vitally important in understanding the dynamics of Internet traffic and in appropriately dimensioning network resources for network and systems management. The vast majority of modelling techniques developed for volume-based traffic profiling (based on packet and/byte counts) imply the statistical assumptions of stationarity, Gaussianity and linearity, which are often taken for granted without being explicitly validated. In this paper, we demonstrate that such properties are often not applicable due to the high fluctuations in Internet traffic, and should therefore be validated first before they are assumed. We employ Time-Frequency (TF) representations and the Hinich algorithms for validating these three modelling assumptions on real backbone and edge network traces. We show by conducting a passive, offline statistical analysis on real operational network traffic traces from both backbone and edge links that link traffic is extremely dynamic irrespective of the level of aggregation and that model characteristics vary. Subsequently, we propose the use of a representative of higher order spectra, the bispectrum, to act as a particularly suitable method for volume-based traffic profiling due to its ability to adapt to different underlying statistical assumptions, as opposed to ARIMA timeseries models that have been typically used in the literature. We demonstrate that the bispectrum, a signal processing tool that has so far been used in the area of image processing and acoustic signals, can be exploited to accurately characterise traffic volumes per transport protocol, and can therefore contribute to fine-grained network operations tasks such as application classification and anomaly detection.

© 2018 The Author(s). Published by Elsevier B.V.

This is an open access article under the CC BY license. (<http://creativecommons.org/licenses/by/4.0/>)

1. Introduction

Understanding the underlying transport-layer traffic behaviour of backbone and edge networks is vital for traffic engineering tasks such as link capacity planning, traffic classification, and anomaly detection [1–5]. Traffic characterisation is typically addressed through statistical analysis of individual link(s) and network-wide traffic volume properties such as counts of bytes and packets [1,2,6,7], as well as by analyzing the distributional behaviour of particular packet header fields [1,4,5,8–10].

In the literature, numerous statistical and signal processing techniques have been proposed to construct traffic models. A number of studies attempt to determine the long-term network-wide

behaviour based on past measurements using wavelet filters, Holt-Winters forecasting, Analysis of Variance (ANOVA) and Autoregressive Integrated Moving Average (ARIMA) methods (e.g. [6,11,14]). At the same time there are volume-based schemes that characterize network behaviour by optimizing the traffic matrix estimation problem (e.g. [7,12,13]). Having as a basis pre-captured IP packet data and Simple Network Management Protocol (SNMP) counters from Points of Presence (PoPs) aggregated and mapped as Origin-Destination (OD) flows, such methods can derive general models to represent overall network traffic demands and routing. Throughout past literature, the most influential theoretical propositions underpinning the traffic matrix problem have been the network tomography Gaussian models provided by Cao et al. [15], Goldschmidt's deterministic linear models derived by Linear Programming (LP) methods [16], and Zhang et al.'s optimized gravity model [12]. Other notable work includes the linear state space modelling [17], and the joint model of Hidden-Markov-Model (HMM) and Gaussian distributions [18].

* Corresponding author.

E-mail addresses: angelos.marnerides@lancaster.ac.uk (A.K. Marnerides), dimitrios.pezaros@glasgow.ac.uk (D.P. Pezaros), d.hutchison@lancaster.ac.uk (D. Hutchison).

Apart from the traffic matrix approach, there have been studies mapping application-related traffic demands to the transport-layer traffic intensity observed on backbone links. For instance, one of the first significant studies by Fieldmann et al., employed a wavelet-based approach to show that scaling properties of Wide Area Network (WAN) traffic are linked to the high activity of Web flows [19]. Per-application traffic characterisation studies have also used signal processing, graph-theoretic and machine-learning approaches to classify application layer activity [5,8,20]. Within anomaly detection, the work presented in [21] proposed a non-Gaussian model for characterizing the volume of unidirectional flows traversing a backbone link. In contrast to the studies mentioned above, the authors in [21] indicate that the strongly asymmetric traffic profile at a transit or backbone link forbids the employment of the well-known OD approach as well as the analysis of bidirectional flows. Moreover, the authors in [22] demonstrate that in order to map the aggregate Internet traffic volume under a Gaussian distribution several penalties need to be employed on a given model that surely do not capture the actual dynamics persisting in the examined measurements.

1.1. Problem definition

Most of the above parametric solutions assume a complete knowledge of the probability distribution of the traffic volume, either on a network-wide, single PoP, or link volume traffic characterisation. Hence, they have incorporated mathematical models based on the *de facto* statistical assumptions of *stationarity*, *Gaussianity* and *linearity*. In general, these statistical properties determine whether the traffic volume may be represented as a stochastic process where 1st- and 2nd-order moments such as mean and variance do not change throughout time (i.e., are *stationary*). In parallel, the traffic volume is assumed to be modeled under a *Gaussian* fit that follows a normal distribution. Moreover, if the observations that compose the traffic volume (in our case byte/packet counts of a flow) have a linear relationship with the preceding or following observations then it can also be represented as a linear state model since it complies with the *linearity* assumption.

Nevertheless, the majority of studies within the realms of Internet traffic characterisation assume these three properties without rigorous validation, whereas others as in [6,13,23] use 2nd-order statistics (i.e. mean, variance, autocorrelation sequence, power spectrum) to validate these assumptions. However, 2nd-order statistics are problematic in validating timeseries properties such as stationarity, linearity and Gaussianity. As explicitly demonstrated in signal processing studies presented in [24] and [25], 2nd-order statistics suppress phase characteristics such as the phase magnitude thus they are unable to capture phase transition peaks. In networking terms, phase transition peaks relate with the adequate capturing of the exact timing and duration of traffic fluctuations occurring in a network [26]. Consequently and as we show in this paper (Section 5.4), any further modelling based on the three assumptions, without a pre-validation process, leads to questionable accuracy and hence this naturally leads to inaccurate interpretation of network traffic dynamics.

1.2. Contributions

In this paper, we focus on the rigorous validation of the statistical properties of *stationarity*, *Gaussianity* and *linearity* using 3rd order statistics, which have been used in the past to model queuing performance [16] and packet interarrival time processes [5]. Here, we use 3rd order statistics for volume-based traffic modelling based on byte and packet counts in unidirectional traffic flows on an aggregate as well as per-protocol basis. Our work sheds new light in the domain of Internet traffic characterisation

through demonstrating the applicability of techniques that are typically used in other domains (e.g., image processing and acoustic signals processing) such as Time-Frequency (TF) representations and 3rd order statistics estimated by the Hinich algorithms and the bispectrum [24,25,27].

We highlight the main contributions from our study:

(1) We first empirically show that the three often-used, *de facto* assumptions of stationarity, Gaussianity, and linearity should be rigorously validated before being applied, as they do not hold universally. The empirical analysis exhibited in this work relies upon real backbone and edge traffic traces and the employment of higher order statistics by the Hinich algorithms. In particular we have found that:

- The non-linear behaviour of the instantaneous frequency and group delay, in all of our datasets on a protocol-specific and aggregate level in both packet and bytes, indicates a highly non-stationary persona in the examined traffic traces.
- All of our datasets on a transport-layer protocol, aggregate volume basis either on a byte or packet-based analysis, indicated to follow non-Gaussian properties.
- In the majority of cases and even in a small temporal interval in the same day (e.g., 30 min.), different transport-layer protocols do not comply with the same modelling assumptions (e.g., Gaussianity, linearity). Therefore their independent, per-protocol analysis is crucial for several traffic characterisation domains such as anomaly detection.
- Even in the process of independent protocol analysis, packets and bytes should be separately profiled since their distributional behaviours exhibit opposing and varying statistical properties with respect to linearity.
- We observe that traffic on the transport layer has many sudden changes within a small time period (e.g., 15 min.) and we also evidently show that the linearity assumptions often fail. Thus, it is essential to reconsider their validity at the pre-modelling stage, since they constitute the underlying basis for selecting a correct traffic characterisation scheme.
- With the use of measurements that have a 10 year gap between them (e.g., 2006–2016) we show that the Internet traffic volume on backbone links from an aggregate or protocol-specific viewpoint still holds the same highly non-stationary and non-Gaussian properties where linearity is evident in some instances.

(2) We show that methods derived by 2nd-order statistics, commonly used for traffic profiling such as ARIMA models and the power spectrum (e.g. as in [6,7,10,14,15,19,21]), often fail to detect traffic fluctuations throughout the observational time frame due to their explicit dependency on assuming the stationary and linearity property for the traffic volume. Apart from not being able to capture phase information, such techniques are also unable to accurately localize frequency fluctuations on the time domain. As we show in this work, such sudden frequency changes undetected by 2nd order statistics are mapped to sudden traffic changes triggered by adverse or malicious events of which network operators need to be aware.

(3) In order to overcome the limitations of the 2nd-order techniques, we propose and validate the use of higher order spectra with the use of the bispectrum [24,25]. The bispectrum is a tool for examining a timeseries' 3rd-order statistical properties which captures phase information and therefore can adequately map traffic fluctuations on the time-frequency plane. We compare the bispectrum with its commonly used 2nd-order counterpart, the power spectrum [24,25], and show that our proposed bispectrum offers a much better resolution on the TF plane leading to higher accuracy in characterizing the behaviour of aggregate traffic volume.

Table 1
Captured Operational Traces from EQUCH, WIDE & Keio.

Set	Date	Duration	Link Type	Packets	Bytes	Flows/min
<i>EQUCH</i>	Mar. 17 2016	60 min	Backbone	1.1G	677G	495 K
<i>WIDE</i>	Mar. 03 2006	55 min	Backbone	32M	14G	63 K
<i>Keio-I</i>	Aug. 06 2006	30 min	Edge	27M	16G	32 K
<i>Keio-II</i>	Aug. 10 2006	30 min	Edge	25M	16G	19 K

(4) We demonstrate the bispectrum's applicability for practical network capacity planning purposes, because of its ability to adapt to different underlying statistical assumptions, which makes it ideal for link traffic volume peak analysis. On the other hand, we show that traditional methodologies such as the ARIMA timeseries (e.g. as in [6,11,13,14,23]) may not be in a position to precisely detect such peaks affecting the overall capacity planning process, due to their explicit dependency on assuming the stationary and linear property for the traffic volume.

The remainder of this paper is structured as follows: Section 2 describes the traffic traces used for our analysis while Section 3 describes the theoretical background and the results obtained from the stationarity test. Section 4 presents the theory behind the Hinich algorithms that underpin the validation of linearity and Gaussianity, and presents the results. Section 5 compares 2nd- and 3rd-order statistics when applied on aggregate traffic volume, and discusses the practical benefits offered by the proposed bispectrum method, especially in the cases of fluctuating traffic demands. It also elaborates on the benefits of per-protocol independent modelling and shows the advantages offered by the bispectrum. Finally Section 6 summarizes the contributions of this work and concludes the paper.

2. Data description

This section is dedicated to presenting and describing the data used within our experimentation. Our analysis is based on unidirectional traffic flows extracted from four anonymized packet payload traces collected in the US and Japan, as shown in Table 1. The EQUCH dataset is a subset of a larger CAIDA dataset¹ collected in 2016 at the Equinix² datacenter in Chicago, IL, that is connected to a 10 Gb/s backbone link of a Tier-1 ISP between Chicago, IL and Seattle, WA. The WIDE trace was collected on a 100 Mb/s US-Japan Trans-Pacific backbone link carrying commodity traffic for WIDE member organizations³. The Keio traces were captured on an 1 Gb/s Ethernet link from the Japanese academic network of the Keio University's Shonan-Fujisawa campus. With the use of CAIDA's CoralReef tool⁴ we grouped the captured packets into their corresponding network flows and then computed per unidirectional transport flow (i.e. TCP, UDP, ICMP) volume statistics (e.g. counts of bytes, counts of packets).

From an application point of view, the EQUCH trace was dominated mainly by WWW (i.e. HTTP/HTTPS) traffic flows but also some considerable amount of unclassified TCP and UDP flows that we speculate were related to game traffic. In addition, the EQUCH trace is also comprised of DNS, RTMP, IPSEC, SSH and also game-related traffic with the QUAKE protocol. On the other hand and as presented in [5], the -10 years older- WIDE trace is mostly comprised of DNS flows, being followed by WWW (i.e. HTTP/HTTPS), FTP as well as unclassified traffic (by the payload-based `crl_pay`

classifier [5]) and general network operation traffic such as Net-Bios, NTP, SNMP and Spamassassin. The Keio-I trace includes FTP, WWW, DNS, and Email traffic. In addition, some scanning attack traffic and a percentage of unknown traffic flows were observed. The Keio-II trace contains DNS, FTP transfers, WWW and streaming media traffic related to applications such as Realplayer, Windows Media Player and Quicktime, as well as P2P flows (e.g. BitTorrent). Though our analysis in this paper is primarily based on decomposing transport traffic into byte and packet counts of TCP, UDP, and ICMP traffic, still the more fine-grained application-based view helped us perform drill-down analyses on observed transport protocol behaviour and identify causal applications or attacks (Section 5.5).

3. Stationarity test

This section presents the experimentation conducted for validating the stationarity assumption in our datasets. It firstly presents via Section 3.1 the signal-oriented principles of instantaneous frequency and group delay, which are used as the metrics for validating stationarity on the byte/packet count timeseries. Subsequently, Section 3.2 demonstrates and discusses the results obtained by our stationarity analysis.

3.1. Methodology

Traditionally in statistics, a stochastic process is said to be stationary if there exists time invariance between observations. From a signal processing point of view, a signal is considered to be *wide-sense stationary* if its decomposition can be expressed as a discrete sum of sinusoids (i.e. as a sum of elements that have constant instantaneous amplitude and instantaneous frequency) [25]. Apart from keeping a constant mean and variance, a wide-sense stationary signal $g(t)$ should also be described by an autocorrelation function (i.e. ACF) $E[g(t_1)g^*(t_2)]$ that only depends on the time difference $t_2 - t_1$. In simple words the ACF relies on a single time lag and does not change with the time at which the function was calculated.

If we wish to non-parametrically characterise a wide-sense stationary signal on the Time-Frequency (TF) plane, it is firstly required to be in its analytical complex form. Under this form, it is expected that the sum of elements composing that signal should keep a constant amplitude and instantaneous frequency respectively as depicted in Fig. 2. In case the signal does not follow any of these constraints, it is considered as non-stationary. Let $g(t)$ denote a signal representing our byte/packet count timeseries and its complex form $G_a(t)$ derived after a Hilbert transformation where $G_a(t) = g(t) + iH[g(t)]$ [25]. Its absolute value $|G_a(t)|$ gives the magnitude of change in the signal (amplitude) in bytes/packets for a given time t . Since the instantaneous peak in the signal is known, through the instantaneous amplitude we can use Eq. (1) to get a measure of the instantaneous frequency $f(t)$:

$$f(t) = \frac{1}{2\pi} \frac{d \arg G_a(t)}{dt} \quad (1)$$

In our case, $f(t)$ denotes the amplitude of frequency we observe in one count of a packet/byte arrival at a particular time t . We also

¹ CAIDA anonymized Internet Traces 2016: http://www.caida.org/data/passive/passive_2016_dataset.xml.

² Equinix: <http://www.equinix.com/>.

³ WIDE MAWI Working Group: <http://mawi.wide.ad.jp/>.

⁴ CAIDA CoralReef Software Suite, available at: <https://www.caida.org/tools/measurement/coralreef/>.

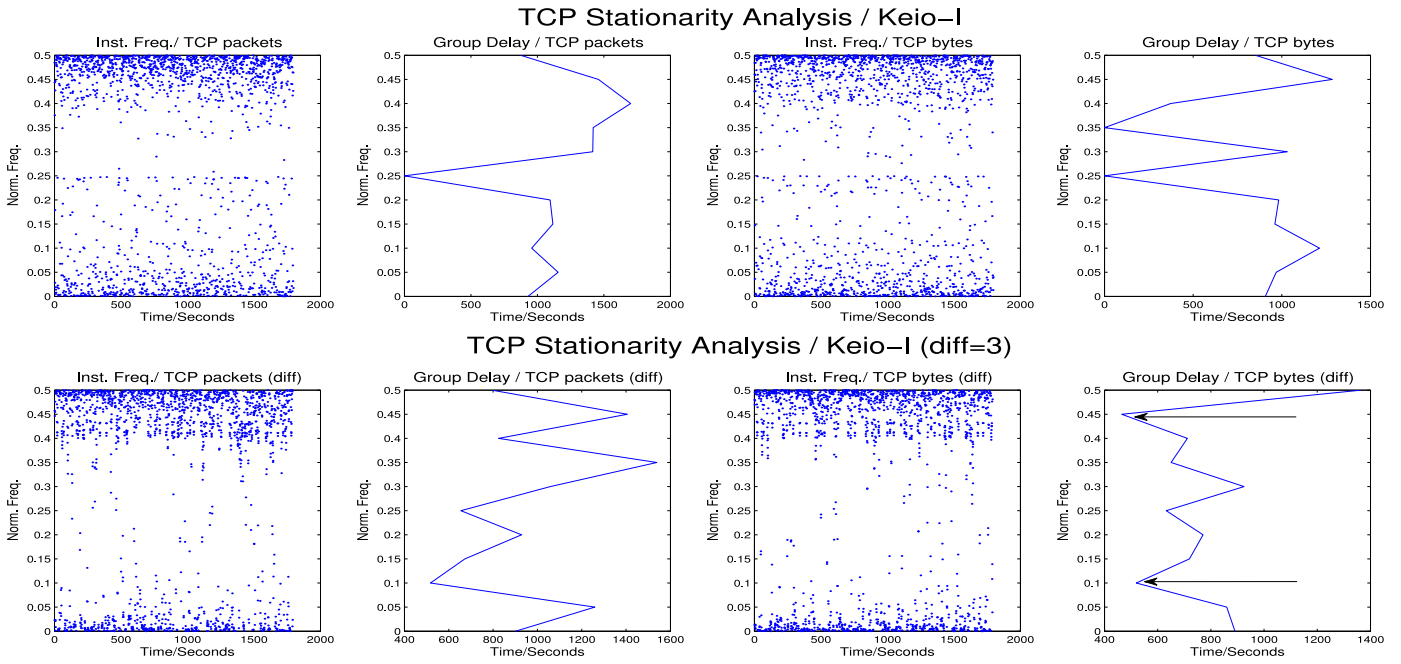


Fig. 1. TCP Keio-I stationarity analysis before (top 4 plots) and after 3rd order differentiation (bottom 4 plots). The black arrows show an example of simultaneous observation in the differenced series.

use the group delay which shows the local time behaviour with respect to the frequency function (i.e. time distortion caused by the signal's instantaneous frequency). The group delay $t_G(\nu)$ is computed by solving Eq. (1) with respect to the Fourier transform of $G_a(t)$, $F_a(\nu)$:

$$t_G(\nu) = \frac{1}{2\pi} \frac{d \arg F_a(\nu)}{d\nu} \quad (2)$$

The definitions provided by the above equations are vital for our validation regarding the stationary behaviour in the byte/packet count timeseries. Hence, we are particularly interested in identifying whether they keep a constant and linear behaviour. As shown next, we observe the characteristics of these metrics and conclude on the inexistence of stationarity.

3.2. Stationarity test: results

This section presents the results obtained from the mathematical equations illustrated in Section 3.1 for validating stationarity. Under the assumption of our traffic signal being stationary and given the definitions of instantaneous frequency and group delay in Section 3.1, we anticipate that each independent byte/packet count timeseries would exhibit unique instantaneous frequency and group delay values. This intuition relates to the assumption that each protocol would behave as a mono-component signal either from the packets or bytes analysis. A mono-component signal behaviour relies on the fact that the counts of both packets and bytes numerically differ and their counts enforce modulated amplitude with complex sinusoidal signals under different time periods. Therefore resulting with a distinct time instant with a single dominant amplitude.

In simple networking terms, such a single dominant amplitude is translated as the case where the volume aggregate of protocol flows (e.g., TCP flows) on a given time instant indicates only a single peak initiated by a particular byte/packet count of a single flow, not by multiple flows. Thus, we cannot have multiple flows indicating a peak on the aggregate but only one which holds the highest byte/packet count. In order to assume a stationary behaviour of our mono-component signal, we expected that each unique frequency

which relates with a single amplitude (i.e., peak initiated by a single, most dominant flow) alongside the rest of the subsequent estimated frequencies would have a constant and close-to-linear evolution in time. Similarly, group delay should present a normalized frequency close to linear with respect to time.

However, as depicted by Fig. 1, none of the anticipated outcomes of constant or linear characteristics for group delay and instantaneous frequency is observed, thus implying a non-stationary behaviour in our datasets. Due to the fact that all the results in all the three tested protocols lead to the same conclusions in all our traffic traces, we restrict this section on presenting results obtained from the Keio-I trace and particularly the stationarity analysis on TCP. Fig. 1 presents the evaluation conducted on Keio-I for TCP from the perspectives of byte and packet counts. The four upper plots in 1 show the behaviour of the instantaneous frequency and group delay on TCP packets/bytes when the initial packet/byte timeseries are not differenced whereas the bottom plots illustrate the same analysis when the packet byte timeseries are differenced up to the third order. At first glance, it is fairly obvious that in all cases, multiple instantaneous frequencies exist within the same time instant. In addition, the group delay outcomes indicate that every signal suffers from time distortion and phase delay as shown by the unstable, non-linear graphs produced. Naturally, the time distortion indicates that the exact capturing of the instantaneous frequency which is related with the instant amplitude of the signal is not correctly adjusted on the TF plane, thus time shifts of individual components (i.e., byte/packets counts of a single flow) are not properly calibrated. These outcomes contradict the assumption that the datasets are of a mono-component nature, since there are different instantaneous frequencies (i.e., multiple amplitudes composed by various flows) and time distortions in the same time instant; therefore the signals have a multi-component nature.

In order to remove some of the distortion and possibly get closer to stationary characteristics we have used a common data analysis technique by differencing our timeseries (e.g., as in [6,14,16]). We have examined our signals up to the 3rd order of differentiation but, as the four bottom plots of Fig. 1 show, the general conclusions with respect to the non-stationary persona of

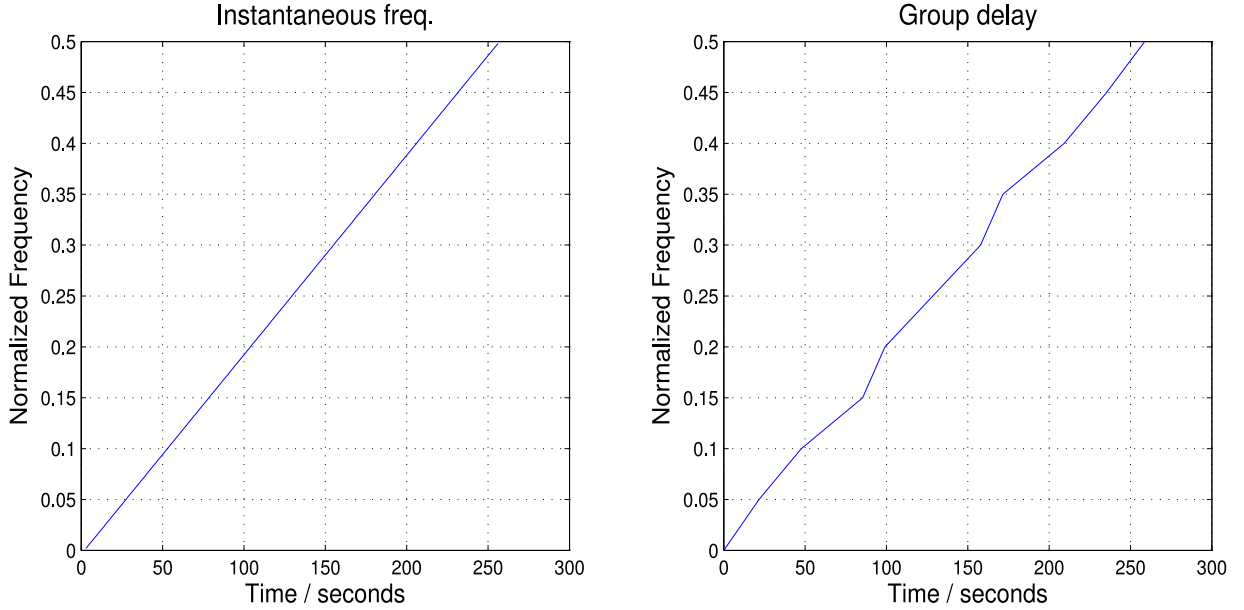


Fig. 2. The ideal shape of instantaneous frequency and group delay representations when a mono-component signal is considered as stationary.

the byte and packet timeseries remain the same. A comparison between the results gained for the non-differenced series (four top plots of Fig. 1) with those attained for the differenced series shows an insignificant reduction of distortion, but still there are random simultaneous frequencies on both byte and packet counts. The outcomes of the four bottom plots in Fig. 1 clearly show that in certain cases the time and frequency distortion is eliminated to a certain extent, but still all signals are structured by several formants. For instance, the group delay graph for the differenced series of TCP bytes illustrates that at 500 seconds there are two different instantaneous frequencies for the TCP byte count signal (indicated with arrows - bottom right plot). Hence, the initial conclusion with respect to the non-stationary, multi-component nature of our protocols is still valid even after differentiating the datasets.

4. Linearity & Gaussianity tests

This section consists of two main parts. Section 4.1 introduces the methodology employed for validating Gaussianity and linearity. In particular, Section 4.1 and 4.2 elaborate upon the concepts of cumulants, bispectrum and bicoherence since they constitute the basic elements of the Hinich algorithms [27], which are used for assessing Gaussianity and linearity in our datasets. Section 4.3 illustrates and discusses the results obtained. Through our analysis, we validate (i) whether traffic flows on the aggregate and on a protocol-specific basis can be adequately fitted under a Gaussian fit and (ii) if the frequency of flow occurrences on the time domain exhibits linear dependencies. We show that each transport layer protocol independently exhibits different statistical characteristics with respect to Gaussianity and linearity within reasonably small time bins.

4.1. Methodology

The identification of linear and Gaussian behaviour has been formulated by employing computational algorithms provided by Hinich [27]. The algorithms are suitable for accurately validating whether a signal's timeseries has linear and Gaussian characteristics. This capability exists due to their direct assessment of a timeseries' 3rd order statistics, serving as the means for profiling phase properties which are not detected by 2nd order statis-

tics (e.g., mean, variance, autocorrelation sequence) [24,25,27]. The Hinich algorithms mainly expose the 3rd order cumulant characteristics of a signal as indicated by the bispectrum and bicoherence values [24,25,27]. In this subsection we introduce the concepts of cumulants, bispectrum and bicoherence, as well as some of their composite statistical features and how these are mapped to our experimental analysis.

4.1.1. From moments to cumulants and polyspectra

Moments give a quantitative measure of the linear combination of points in the distribution of a random process. On the other hand, cumulants denote the non-linear combinations of points within the distribution for a random process $g(t)$. Traditionally, the estimation of the well-known power spectrum which defines the spectral density of a process is derived from the Fourier Transform (FT) of the 2nd order moment sequence (i.e. autocorrelation sequence). Due to the elaboration on Kolmogorov's moment definitions provided by Rosenblatt [24], the power spectrum (also known as the 2nd order polyspectrum) can be defined in terms of cumulants instead of moments. Subsequently, we can express the power spectrum $S(\omega_1)$ with respect to the 2nd order cumulant sequence $c_2(\tau_1)$ as:

$$S(\omega_1) = \sum_{\tau_1=-\infty}^{+\infty} c_2(\tau_1) e^{-j(\omega_1 \tau_1)} \quad (3)$$

On the other hand, the topic of interest in the Hinich algorithms is the estimation of the bispectrum which is the 3rd order polyspectrum. According to [24], higher spectra of order $N \geq 3$ are strictly expressed by cumulants. The bispectrum provides a dual-frequency representation on the time-frequency (TF) plane in contrast to the one-dimensional interpretation of the power spectrum, and as shown by Eq. (4), it is considering the 3rd order cumulant sequence $c_3(\tau_1, \tau_2)$ as its basic tuning function:

$$B(\omega_1, \omega_2) = \sum_{\tau_1=-\infty}^{+\infty} \sum_{\tau_2=-\infty}^{+\infty} c_3(\tau_1, \tau_2) e^{-j(\omega_1 \tau_1 + \omega_2 \tau_2)} \quad (4)$$

Work in [24] and [25] suggest that the bispectrum provides a much finer level of granularity for detecting and further characterizing a non-linear signal than the power spectrum. The reason is that

the power spectrum directly depends on the autocorrelation sequence, whereas the bispectrum deals with the 3rd order cumulant sequence. In addition, the bispectrum can extract information related to the minimum phase shifts that take place on a non-linear multi-component signal. As we show in Section 4.3 this is the case for Internet traffic and particularly on per-protocol counts of bytes and packets where most of them exhibit different phase properties and in many cases, a non-linear property.

Based on the aforementioned characteristics, we propose to use the bispectrum as the basic means of characterizing our traffic streams in order to verify the assumptions of linearity and Gaussianity. Due to the fact that we could not obtain a crisp understanding regarding the probability distribution function of our observed timeseries we apply a *non-parametric* estimation of the bispectrum [28]. Used alongside bicoherence that we explain next, the bispectrum also reveals other distributional properties such as the skewness of our dataset. Specifically, the Gaussianity test is commonly mentioned as the “zero-skewness test” since theoretically a dataset exhibiting zero skewness is considered to be Gaussian [28]. Although skewness is derived from the 3rd order cumulant sequence (as the bispectrum and bicoherence are), zero valued skewness is only valid if it satisfies the requirement that our dataset shows zero mean, which in our case is not. Therefore, we consider it more appropriate to use the bispectrum and its squared normalized version (i.e. bicoherence) which do not require the zero mean as a prerequisite.

Bicoherence is a useful statistical tool derived from the definition of coherence. Coherence describes frequency-domain correlations for a given signal, and it is also established as a means for measuring linear dependencies of moments. On the contrary, bicoherence is used to detect quadratic non-linearities on the time-frequency (TF) plane as well as quadratic phase coupling, denoted as the squared normalized version of the bispectrum as follows:

$$b_{k_B(\omega_1, \omega_2)} = \frac{B(\omega_1, \omega_2)}{\sqrt{S(\omega_1 + \omega_2)S(\omega_1)S(\omega_2)}} \quad (5)$$

$S(\omega_1)$ and $S(\omega_2)$ are the power spectrums of two independent Fourier components, and $S(\omega_1 + \omega_2)$ is the power spectrum of a composite 3rd signal defined by the two independent Fourier signals ω_1 and ω_2 . Eq. (5) is considered a crucial element of the Hinich algorithms. It is actually the basic means for estimating the distance of cumulants from each other which in our process represents the counts of bytes and packets for each flow.

4.2. Hinich algorithms

Firstly introduced in [11], the Hinich⁵ algorithms are a set of statistical hypothesis tests to detect non-linear and further Gaussian or non-Gaussian characteristics on a given random stochastic process $g(t)$. As already mentioned above, the process $g(t)$ in our case is the count of packets or bytes in discrete time bins of length n , denoted as $T = \tau_1, \tau_2, \dots, \tau_n$. The following subsections introduce the basic equations employed by the Gaussianity and linearity tests.

⁵ Naturally the concepts invoked by the Hinich algorithm and the bispectrum involve non-parametric signal processing schemes that in practice consider the random traffic behaviour where its probability density function either on a packet or byte perspective cannot be concretely defined as commonly assumed in the literature. Albeit that the Hinich algorithms have been challenged due to their dependency on Gaussian asymptotics as well as their suboptimal smoothing in the bispectral domain [29] we still argue in fair of their validity since the opposing suggestions do not provide constructive comparisons on real datasets but they rather equate synthetic timeseries of oscillatory random processes in extremely small time intervals [29]. Thus, given the outcomes of several Internet traffic characterisation schemes we can characterize the traffic volume timeseries as the observable state of a random process but on the other hand we do not hold all the properties (e.g. frequency stability) that are explicit to oscillatory random processes.

4.2.1. Gaussianity test

This test focuses on the bicoherence value as the result of the normalized bispectrum estimate. The Gaussianity (also known as zero-skewness) assumption according to Hinich is considered the case where the estimated bicoherence value $E[b_{k_B(\omega_1, \omega_2)}]$ as well as the skewness equal to zero. Since in reality we can never have a flat, zero-bicoherence (due to noise), we take the mean bicoherence value which in practice represents a quantitative Gaussianity metric [24,25,27]. By using the definition of Eq. (5) we can calculate the mean bicoherence power estimate k with:

$$k = \sum |b_{k_B(\omega_1, \omega_2)}|^2 \quad (6)$$

Work demonstrated in [10] and [3] employ some comparisons of the k value in order to establish a concrete conclusion on whether the dataset under test actually follows a Gaussian distribution or not. As they suggest, the computed value of k is χ^2 distributed (chi-squared distributed) with a Fast Fourier Transform (FFT) length function for approximating the number of freedom degrees to the closest Gaussian fit. Therefore, in case the estimated k indicates that our timeseries has less than or equal to 10 freedom degrees, then we conclude that the dataset follows a Gaussian distribution. In parallel with k , the Zero Skewness Probability (ZSP) of false alarm ϑ is computed. ϑ illustrates the probability of the newly approximated value being much larger than the initial k estimate. When ϑ is small, we can reject the zero-skewness (i.e. Gaussian) hypothesis. In our tests we use $0.2 \leq \vartheta \leq 1$ as a valid range for approving Gaussianity, based on findings from work in [24,25,27,30].

4.2.2. Linearity test

The linearity assumption holds in case where the bispectrum of a real process $g(t)$:

$$B(\omega_1, \omega_2) \neq 0 \\ = C, \forall \omega_1, \omega_2 \quad (7)$$

where C is a constant value for all ω_1 and ω_2 . As suggested by Swami [28], it is essential to approximate values of a sample interquartile range under a new bispectral estimate $\Psi(\omega_1, \omega_2)$ derived by $B(\omega_1, \omega_2)$:

$$\Psi(\omega_1, \omega_2) = \frac{1}{\sqrt{M^{1-2\rho}}} B(\omega_1, \omega_2) \quad (8)$$

where M is the resulting boxcar window length, after rounding the FFT length with a resolution parameter ρ . The intuition for calculating the interquartile range is to compare it with a theoretical interquartile range $Y(\omega_1, \omega_2)$ which similarly to $\Psi(\omega_1, \omega_2)$ is chi-squared with a non-centrality parameter η defined as:

$$\eta = (2M^{2\rho-1})c_3(\tau_1, \tau_2) \quad (9)$$

In our experimentation we keep the values of a FFT length of 128 and a resolution parameter ρ of 0.51 since higher

resolution parameter values would involve drawbacks with respect to frequency resolution [28]. The main step is to compare $\Psi(\omega_1, \omega_2)$ with $Y(\omega_1, \omega_2)$ and if the difference is higher than the limits provided by [27,28] (e.g., >10), then we reject the linearity hypothesis.

4.3. Linearity & Gaussianity analysis: results

This subsection examines the validity of the linearity and Gaussianity assumptions on our datasets, using the Hinich algorithms. Tables 2 and 3 demonstrate the results obtained for all our datasets on an aggregate and transport protocol-specific perspective. In general, all examined our datasets either on a protocol or aggregate-level analysis exhibited zero values on the Zero Skewness Probability (ZSP - ϑ) from both a bytes and packets perspective. Therefore, all the timeseries examined promote a non-Gaussian property. This

Table 2Linearity and Gaussianity analysis results on aggregate packet/byte volume. *L=Linear, NL = Non-Linear, G = Gaussian, NG = non-Gaussian.*

Traffic Dataset	$\Psi(\omega_1, \omega_2)$ estimated	$\Psi(\omega_1, \omega_2)$ theoretical	Gaussianity Metric (<i>k</i>)	Degrees of Freedom (DF)	Conclusion
EQUCH (packets)	39.491	10.682	258.131	12	NL & NG
EQUCH (bytes)	30.271	7.188	312.923	13	NL & NG
WIDE (packets)	10.352	8.1582	187.122	11	L & NG
WIDE (bytes)	29.368	4.331	236.477	12	NL & NG
Keio-I (packets)	18.696	13.498	962.762	14	L & NG
Keio-I (bytes)	22.664	14.988	1444.206	32	L & NG
Keio-II (packets)	21.229	15.865	1461.060	32	L & NG
Keio-II (bytes)	22.390	8.241	1396.049	31	NL & NG

Table 3Per transport protocol packet/byte volume Linearity and Gaussianity analysis. *L=Linear, NL = Non-Linear, G = Gaussian, NG = non-Gaussian.*

Traffic Dataset	$\Psi(\omega_1, \omega_2)$ estimated	$\Psi(\omega_1, \omega_2)$ theoretical	Gaussianity Metric (<i>k</i>)	Degrees of Freedom (DF)	Conclusion
EQUCH TCP packets	22.387	19.263	1378.459	36	L & NG
EQUCH TCP bytes	32.628	20.056	1836.109	42	NL & NG
WIDE TCP packets	11.548	9.279	233.950	13	L & NG
WIDE TCP bytes	23.620	12.099	406.533	17	NL & NG
Keio-I TCP packets	17.752	14.970	907.543	21	L & NG
Keio-I TCP bytes	29.218	23.552	1595.313	38	L & NG
Keio-II TCP packets	18.919	17.006	813.860	19	L & NG
Keio-II TCP bytes	22.076	12.807	1346.127	34	NL & NG
EQUCH UDP packets	24.271	12.117	392.781	15	NL & NG
EQUCH UDP bytes	32.826	19.516	6138.239	72	NL & NG
WIDE UDP packets	9.304	16.066	726.540	19	L & NG
WIDE UDP bytes	39.101	12.997	4304.947	53	NL & NG
Keio-I UDP packets	42.673	14.710	5152.262	55	NL & NG
Keio-I UDP bytes	39.896	15.452	4508.253	51	NL & NG
Keio-II UDP packets	44.190	7.735	5551.960	56	NL & NG
Keio-II UDP bytes	38.038	10.499	4097.493	49	NL & NG
EQUCH ICMP packets	391.598	52.711	7297.150	73	NL & NG
EQUCH ICMP bytes	641.618	48.275	8630.323	75	NL & NG
WIDE ICMP packets	131.699	44.134	5555.263	56	NL & NG
WIDE ICMP bytes	264.411	35.265	3548.293	48	NL & NG
Keio-I ICMP packets	85.826	23.485	1582.503	38	NL & NG
Keio-I ICMP bytes	32.489	10.661	2979.164	40	NL & NG
Keio-II ICMP packets	37.335	46.486	6408.248	72	L & NG
Keio-II ICMP bytes	34.649	39.367	4405.960	53	L & NG

is also justified by the generated Degree of Freedom (DF) values produced for each packet or byte-wise distribution that hold values greater than the acceptable threshold of 10 DFs that we discussed earlier (Section 4.2.1).

As illustrated in Table 2, we observe that the linearity assumption varies in each dataset and is not necessarily related to the particular volume feature (i.e. packet or byte). By looking at the EQUCH dataset, we verify that the traffic volume from both a packet or byte perspective holds non-linear and non-Gaussian properties, thus they both need to be carefully modelled under a representation that severely considers these properties. On the other hand, the aggregate volume properties of the WIDE trace exhibit non-linear and non-Gaussian properties if we strictly consider the bytes distribution whereas the packet distribution demonstrates linear properties and non-Gaussian. Hence, a potential modelling approach for a given traffic characterisation application (e.g. anomaly detection) that requires granular view of these flow features should consider independently the modelling of packets and bytes. Similarly, such a modelling approach should act in the same fashion if it wishes to assess the dynamics of the Keio-II dataset since its aggregate packet volume distribution holds linear and non-Gaussian properties whereas its bytes distribution demonstrates a non-linear and non-Gaussian profile based on the computed statistics. Finally, the Keio-I dataset has linear and non-Gaussian properties on both its packet and byte-wise distribution, thus a potential modelling approach may assume linearity for the overall aggregate traffic volume.

Overall, Table 2 has demonstrated that empirically we do see the linearity assumption to be seen as frequently as the non-

linearity assumption. However, the transport protocol-based volume analysis illustrated by Table 3 indicates that the linearity assumption is not true in most of the examined distributions. It is therefore evident that even when modelling a single protocol, we should consider the counts of bytes and packets independently and analyze them separately. On the other hand and similarly with the aggregate volume analysis, the non-Gaussianity assumption still holds throughout all the datasets from a transport protocol viewpoint.

Given the statistics depicted in Table 3 we can clearly visualise that the transport protocol dynamics in the EQUCH dataset exhibit varying outcomes with respect to linearity on each individual transport protocol (i.e. TCP, UDP, ICMP). For instance, in EQUCH as well as in the WIDE dataset, TCP cannot be modelled in the same way on both packets and bytes since the former behaves linearly and the latter non-linearly. Nonetheless, the UDP protocol in the EQUCH trace can be modelled under the same statistical assumptions since both its packet and byte distributions appear to conform with non-linear and non-Gaussian properties. However, UDP in the WIDE trace has a linear and non-Gaussian behaviour from a packets perspective, thus it should be modelled independently from the packets distribution that demonstrates a non-linear and non-Gaussian property. Furthermore, ICMP in both the EQUCH and the WIDE trace appears to have non-linear and non-Gaussian properties from both a packet or byte-wise perspective. Therefore, both ICMP packets and bytes in this case could be modelled jointly by considering the aforementioned assumptions.

The Keio I/II characterisation presented in Table 3 has produced some significant outcomes related to the statistical behaviour of

each protocol on byte and packet counts. It is observed that even in a small temporal interval in the same day (i.e., 30 mins), TCP and ICMP may not be modeled under the same assumptions. According to the generated statistics assessing linearity and Gaussianity, TCP in Keio-I for both byte and packet counts complies with a linear (i.e., L) and non-Gaussian (i.e., NG) statistical assumption, whereas ICMP was found in both packet and bytes to follow a non-linear (i.e., NL) and NG persona. Even though they can both be modeled within a non-Gaussian model, still a uniform scheme cannot be employed since a Gaussian (or mixed Gaussian) fit to both protocols would not be able to fully capture the varying and unstable changes of the mean and the variance, as implied by the NL profile of ICMP. Nevertheless, the findings indicated in Table 3 illustrate the existence of NG characteristics on the aggregate and all the protocols independently, thus justifying the efforts placed in [6] where traffic profiling was done under a non-Gaussian modelling method. In general, it is also obvious that not all the protocols may be modeled in the same way therefore per-protocol, independent analysis is warranted, as we will present in Section 5.5.

Throughout Keio-I, only ICMP and UDP can be modeled in the same way (i.e., they both fall under the NL & NG assumptions), whereas TCP should be treated under linear but non-Gaussian assumptions. All the three protocols for both packet and byte counts have a zero probability for exhibiting zero skewness. This zero probability is the main strong evidence for concluding that the data do not exhibit Gaussian behaviour. The linearity exhibited by TCP in Keio-I is demonstrated by the small numerical difference (<10) between the theoretical and the estimated interquartile range. On the contrary, we conclude that neither UDP nor ICMP exhibit linear characteristics since the difference between the interquartile ranges gets greater than 10. Identical analysis on Keio-II reinforces the results from Keio-I indicating again that all protocols cannot be modeled together under the same statistical assumptions.

Our experimentation goes a step further in order to empower the argument of the varying outcomes on the fundamental assumptions of linearity and Gaussianity over small timescales. In particular, we assessed the validity of these assumptions in smaller timescales on the EQUCH dataset. In order to achieve this, we initially segmented the EQUCH dataset in smaller samples of 15 minutes (i.e. EQUCH I - IV) and examined the statistical behaviour of their aggregate volume from both a packet and bytes perspective. As evidenced by Fig. 3, for the first 30 minutes in the EQUCH dataset (i.e. EQUCH I, EQUCH II) the aggregate volume on both packets and bytes is considered as linear since the absolute difference of the estimated interquartile range Ψ_{est} with theoretical interquartile range Ψ_{th} is less than 10. However, the statistical assumptions change for the subsequent 15 minutes in EQUCH since the linearity assumption is not valid since the aforementioned interquartile range absolute difference goes beyond the acceptable linearity threshold. Finally, the last 15 minute bin on EQUCH promotes a linear property since the difference gets lower than the threshold. Thus, a potential traffic model that aims to profile traffic dynamics for fine-grained operations (e.g. anomaly detection) would essentially need to reconsider this varying property in order to achieve higher accuracy.

In parallel, the visualization of the Degrees of Freedom (DF) depicted in Fig. 4 indicates that the Gaussianity assumption is not valid throughout the whole 60 minute duration on EQUCH. In fact, all the computed DF values as resulted by chi-squaring the k value using a FFT function, are beyond the acceptable Gaussianity threshold of 10 DFs. Overall, via this small granular traffic analysis we observe the linearity and Gaussianity assumptions failing in small timescales and therefore the need for reconsidering their validity is essential, since they constitute the underlying basis for selecting a correct traffic characterisation scheme.

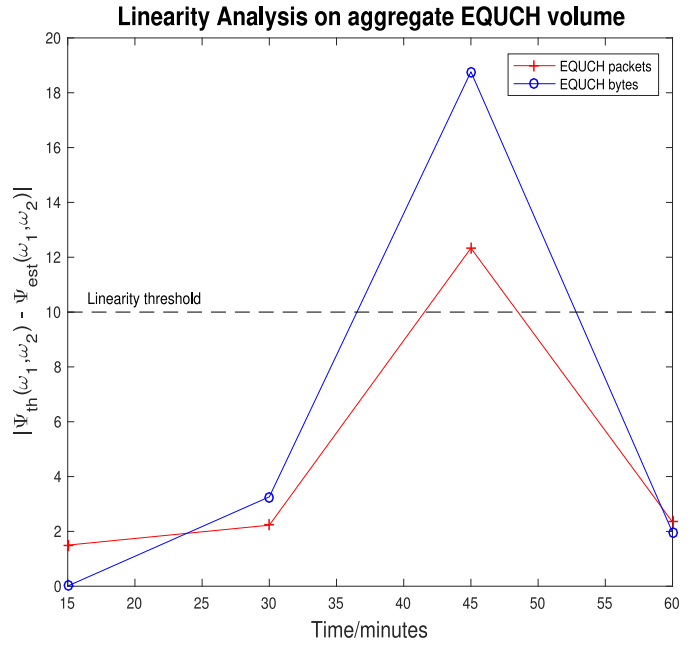


Fig. 3. Linearity analysis on EQUCH on 15 minute timescales.

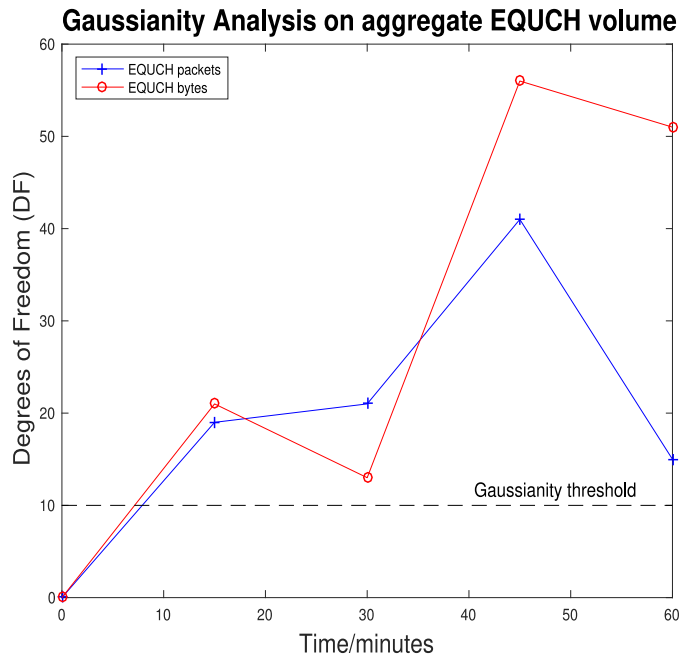


Fig. 4. Gaussianity analysis on EQUCH on 15 minute timescales.

5. Traffic characterisation under higher order spectra

In this section we illustrate the importance of validating statistical assumptions within the context of capacity planning and particularly detection of sudden volume peaks. We compare the performance of our proposed bispectrum as derived out of higher order spectra against the typical ARIMA models employed in several past studies on capacity planning (e.g. [6]). Our choice of the bispectrum is motivated by the fact that (i) packet and byte traffic signals in all of our datasets showed highly non-stationary and non-Gaussian characteristics while linearity was evident in some cases, and (ii) the bispectrum is the most suitable candidate for

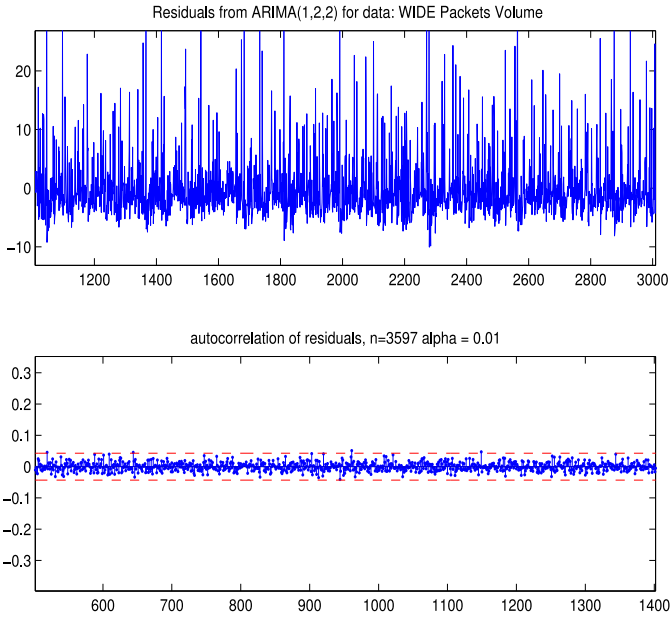


Fig. 5. Residuals (top) and ACF plot (bottom) obtained for WIDE packets. Y-Axis = Residual (top), Autocorrelation Value (bottom), X-Axis = Number of Flows.

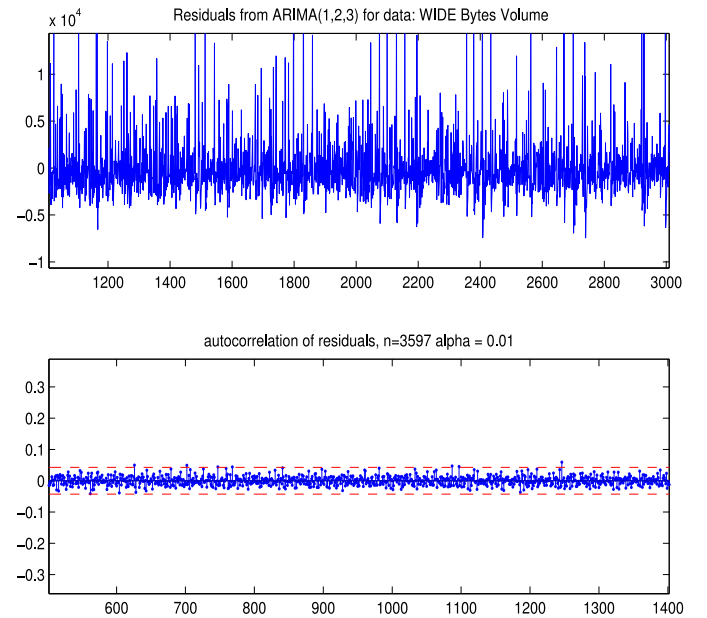


Fig. 6. Residuals (top) and ACF plot (bottom) obtained for WIDE bytes. Y-Axis=Residual (top), Autocorrelation Value (bottom), X-Axis = Number of Flows.

reasonably profiling such signal protocols [24,25], due to its own ability of assessing higher, 3rd-order statistical properties.⁶

5.1. ARIMA models

We have generated several ARIMA models describing the residuals and the autocorrelations that exist in our datasets as well as periodograms that provide a power spectrum representation of the same datasets. The most appropriate fits for our datasets were selected based on the Akaike Information Criterion (AIC) and Akaike's Final Prediction Error (FPE) which are particularly suitable statistical metrics for comparing several linear models⁷. Based on the AIC and FPE values, the most appropriate model to fit the overall counts of bytes in WIDE was a first order Auto-Regressive (AR = 1) second order differentiation ($I = 2$) with three Moving Average steps (MA = 3). On the other hand, the AIC and FPE values obtained for the packet count profiling indicated that the best-fit ARIMA model had a first order Auto-Regressive step (AR = 1) and two Moving Average steps (MA = 2) while the whole series should have been initially differentiated twice ($I = 2$).

As shown in both Figs. 5 and 6, the residual plots (top) generated for both traffic signals cannot determine any trends and the generated autocorrelation (ACF) plots (bottom) empower the speculation of non-stationary and random characteristics. At the same time, both ACF plots illustrate the absence of strong autocorrelation between the observations on the x-axis implying that neither the packet nor byte count signals in WIDE possess self-similar characteristics.

The range of values obtained for the autocorrelation coefficient (i.e. close to 0) justifies the absence of strong or even mild self-similarity for any observation in our signals. Consequently, this latter fact leads to the conclusion that the traffic process as captured

within the WIDE trace exhibits short memory and not *Long-Range Dependence* (LRD). As widely known, LRD measures the memory of a process and is based on the existence of significant correlation between distant events over different time lags within a series. LRD's basic building block is composed by the findings of the ACF and in our case there is not a single observation that holds similar characteristics with itself in a different time lag. In a scenario where LRD would exist we would expect a non-degenerate ACF asymptotically decaying to zero.

Due to the not well-defined patterns and the inexistence of strong autocorrelations depicted by the plots provided by Fig. 5 and Fig. 6, it is clearly evident that the ARIMA model cannot provide a meaningful and interpretable view for the traffic behaviour with respect to its exact evolution on the time domain. We argue that this is due to the nature of the underlying statistical assumptions of linearity and non-stationarity which we validated earlier (Sections 3.2, 4.3). By principle, the AutoRegressive (AR) and Moving Average (MA) processes embodied within an ARIMA model require the linear dependency of observations throughout time and in parallel assumes their stationary behaviour [32].

5.2. Power spectrum vs. bispectrum for traffic surge detection

The resulting power spectrum representations for the WIDE trace as derived by the ARIMA models are shown in Fig. 7. Since the timeseries were initially differentiated up to a 2nd-order by the ARIMA model, we followed the same process for the bispectrum estimation. Fig. 7 shows the generated periodograms and estimated models for ARIMA (1, 2, 3) and ARIMA (1, 2, 2) illustrating that overall byte and packet counts exhibit a close-to-exponential growth with no significant or dominant peaks. This representation implies that there have not been sudden traffic changes in our traces, and that it is likely to have minimal or nonexistent traffic bursts. In addition, even in the case of some small dominant frequency cycles (e.g. on $\lambda = 0.5, 2.5, 3$ for byte and $\lambda = 0.3, 2, 2.3, 3$ for packet counts) that one could be tempted to investigate, the power estimates were not able to capture the exact timing of traffic surge events as seen in the raw datasets. However, by manually inspecting the traces, we did observe certain traffic bursts due to high utilisation of the three transport protocols TCP, UDP and

⁶ Albeit that the primitive estimation of bispectral estimates relies on the weak (but not strong) stationarity assumption, suggestions depicted in [24,25,28] ensured that a fine-grained tuning on the smoothing function within the estimation of the non-parametric indirect bispectrum, can reasonably characterize non-stationary timeseries as well.

⁷ Due to space limitations, we will not explain in detail how the ARIMA modelling, the AIC and the FPE are mathematically defined, rather we refer the interested readers to [31]

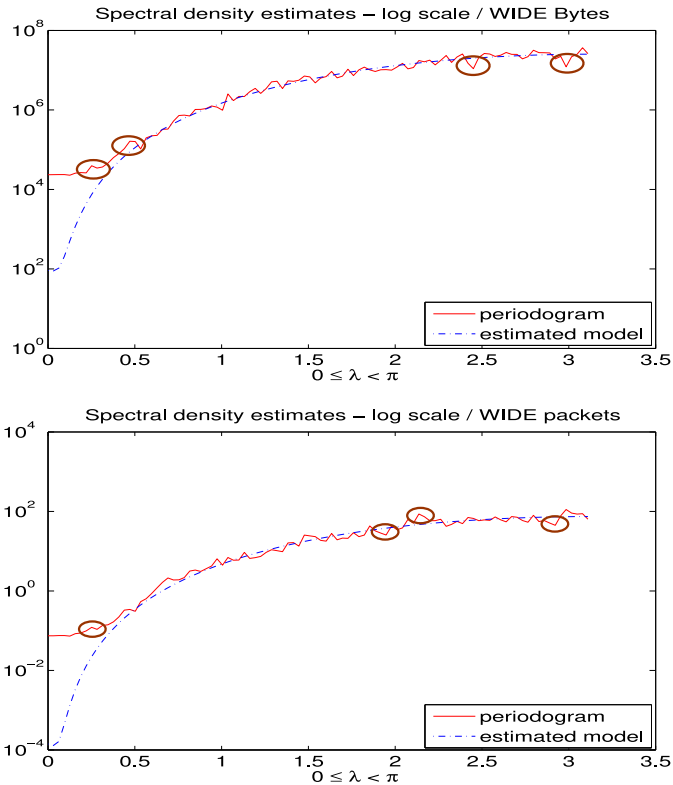


Fig. 7. Periodograms and estimated models produced for WIDE bytes (top) and packets (bottom) based on the ARIMA modelling - traffic surges circled in brown.

ICMP. These high peaks were reasonably detected by the bispectrum which, as shown in Fig. 8, had isolated the phase transitions present in the timeseries.

As illustrated at the top right corner⁸ in both bispectra of Fig. 8, there are several distinct points or regions (coloured in yellow and dark brown those with high amplitudes based on the amplitude legend next to each bispectrum figure) indicating phase transition peaks that explicitly denote sharp traffic surges, which were not detected by the ARIMA model.

Based on manual inspection conducted with the raw dataset, we confirmed that there were some significant traffic demands arising from all the three protocols. As depicted in Fig. 8, the regions surrounding the 2D frequency points in both packet and byte count of (0.26, 0.34) and (0.27, 0.39) exhibit a close-to-synchronised behaviour of increased traffic utilisation. However, as it also happens with the rest of the regions around the points (0.33, 0.28), (0.33, 0.39), (0.39, 0.28) and (0.39, 0.35), we could clearly identify traffic volume surges on both packet and byte counts.

Generally, in contrast to the periodogram output shown in Fig. 7, the generated bispectral estimates were able to localise and further isolate the distinct traffic demands enabling much easier inspection in the raw traces. These results show that the bispectrum-based approach provides a much more accurate TF representation of traffic than the commonly-used spectral representation of a produced ARIMA model via the power spectrum approach.

⁸ Due to the bispectrum's symmetric properties presented in Appendix A, our analysis focuses on the estimates represented on the top right segment of the bispectra images (i.e. on the positive axes) provided in Fig. 8.

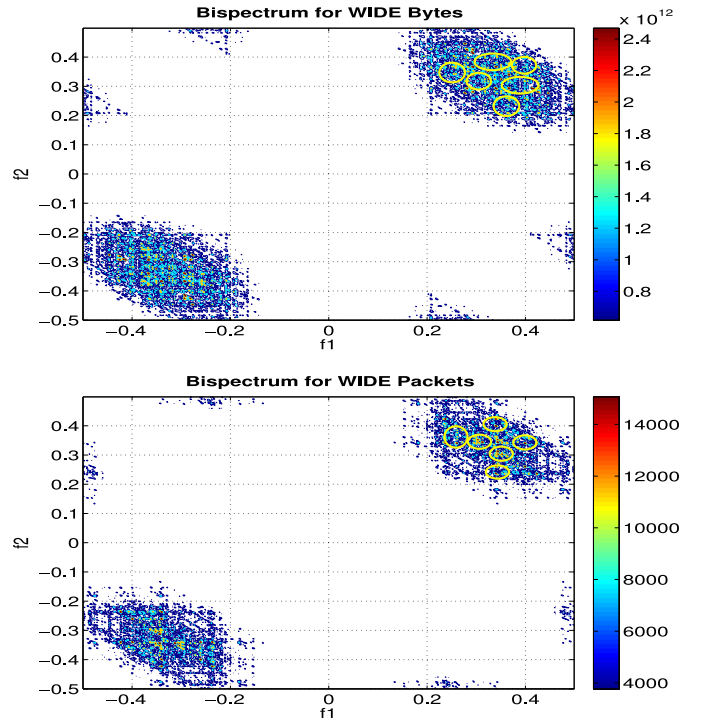


Fig. 8. Bispectra produced for WIDE bytes (top) and packets (bottom) - regions mapped with traffic surges are marked with yellow circles-. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5.3. Traffic peak analysis

The ability to accurately localise high traffic demands on the time dimension is a pre-requisite within the context of capacity planning. Therefore, this subsection demonstrates the applicability of the bispectrum on a traffic peak analysis scenario and we compare its accuracy with that of the ARIMA methodology. We have mapped the signal intensities (i.e. amplitudes) into their corresponding byte-based values (Mbps) and picked the most dominant average traffic volume peaks (i.e., highest averaged mean values) obtained in the WIDE-II bin from the original timeseries signal.

Fig. 9 shows that the bispectral estimates (green plots) outperform the power spectrum approximations (red plots) resulted after the ARIMA modelling on the WIDE trace. The two-dimensional frequency indices provided by the bispectrum's definition (Section 4.1.1) allow the generation of a range of distributions that reasonably match the traffic peaks present in the original signal. Due to the complex nature of the resulting bispectrum we have employed the least squares method in order to extract the best fit distributions compared with the original signal values. As illustrated in Fig. 9, in the case of WIDE-II we found the four best-fit distributions within the overall WIDE-II bispectrum which are by far more accurate than the periodogram. All of the four distributions tend to have a close approximation to the initial highest peak of the original signal ($\approx 95\text{Mbps}$). Despite the fact that most of their subsequent estimations were close to the original signal, it is also evident that particularly the distributions of Fig. 9(a), (b) and (d) are unable to capture the second and third peaks at the 2nd and 3rd minute respectively. In addition, the distributions of Fig. 9(b) and (d) tend not to fully match with the original signal's values on the last minutes of observation (i.e., from the 11th minute and onwards) whereas those of Fig. 9(a) and (c) are reasonably close to these values though not fully matching them. These outcomes are due to the initial addition of the first five cumulants

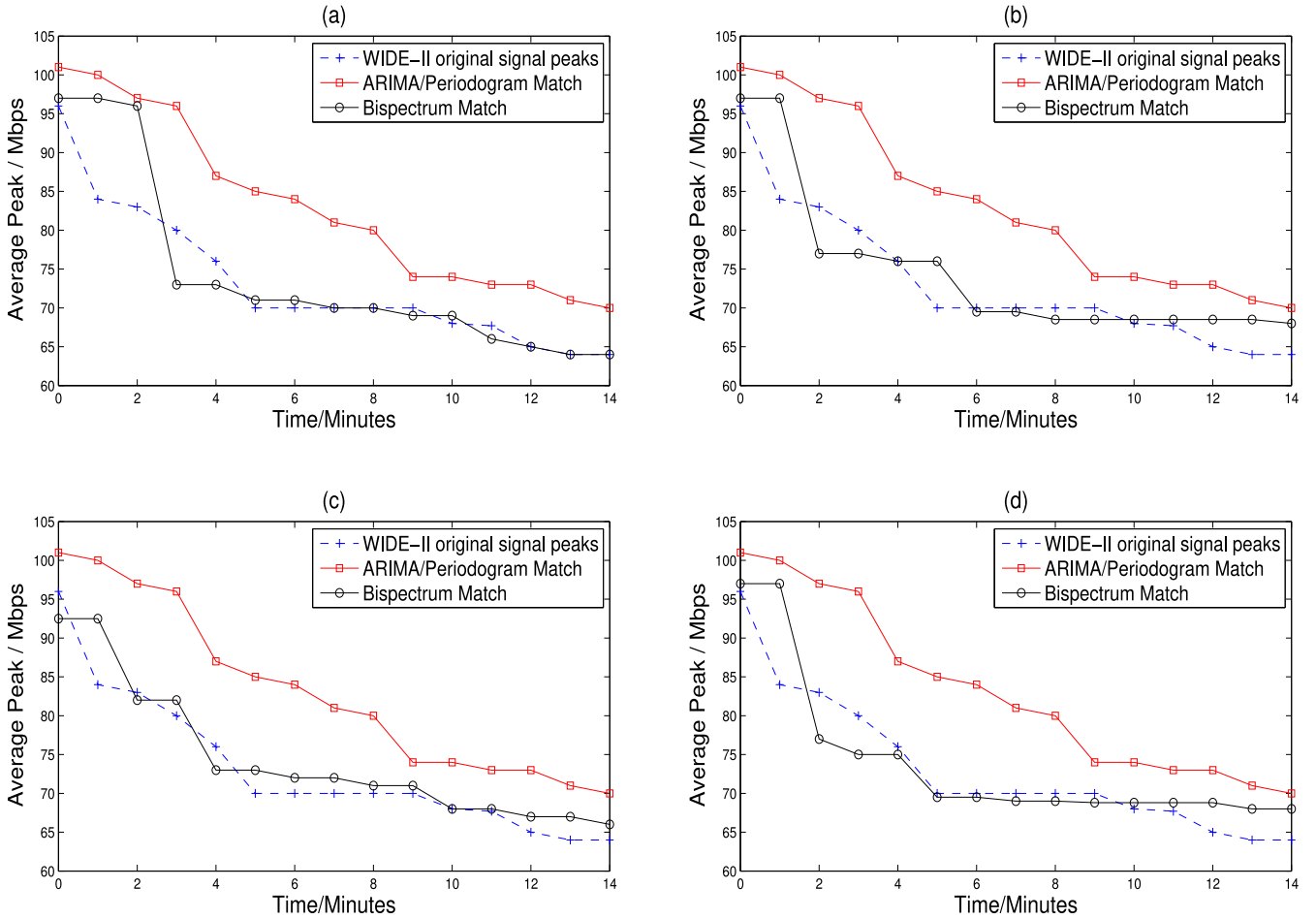


Fig. 9. Aggregate volume peak analysis comparison between the four best distributions offered by the bispectrum and the ARIMA/periodogram method.

within the bispectrum's cumulant sequence $c_3(\tau_1, \tau_2)$ which is the core element of the bispectrum as explained in Section 4.1.1. In some cases this process of computing the 3rd order cumulant sequence involves complex and negative values that cause the dislocation of the resulted bispectral estimates when included within the dual Fourier transformation required by the bispectrum's computation.

Nevertheless, from a capacity planning point of view it is fairly obvious by Fig. 9 that the power spectrum (ARIMA/periodogram) estimations produce more erroneous indications which subsequently lead to wrong conclusions. By considering the fact that the WIDE link holds a maximum capacity of 100Mbps (during the trace collection period), the ARIMA predictions alarm two cases in $t \approx 0$ and $t \approx 1$ where the link is overflowed with instant link utilisation of 102 and 101Mbps respectively. Under a real scenario, the network operator in this case would seriously consider the forecasting provided by the ARIMA model and would regard over-provisioning the infrastructure in the short term (e.g., through redundant links), which would essentially result in unnecessary extra cost for an ISP. On the contrary, the bispectral distributions in all the cases have reasonably matched the majority of peaks and are by far closer to the real valued peaks exposed by the original traffic signal.

5.4. Importance of validating statistical assumptions - ARIMA modelling on volume aggregates

This section presents our aggregate volume analysis on both the Keio traces in order to illustrate the importance of thoroughly validating models' statistical assumptions of stationarity, linearity and

Table 4
Best fit ARIMA models for Keio-I/II.

Trace	Vol.	ARIMA	AIC	FPE
Keio-I	Bytes	(0, 2, 2)	2.211e + 001	4.007e + 009
Keio-I	Packets	(1, 2, 3)	1.576e + 001	6.969e + 006
Keio-II	Bytes	(2, 1, 3)	2.390e + 001	2.386e + 010
Keio-II	Packets	(1, 2, 3)	1.032e + 001	3.038e + 004

Gaussianity. We examine the behaviour of ARIMA on Keio-I and Keio-II and investigate whether a single model may be applied. In addition, we also independently model the counts of packets and bytes on the overall volume including all the transport protocol traffic together.

Similarly to the previous compound volume analysis performed on WIDE, AIC and FPE values are the metrics to determine the suitability of a particular model between the several produced. The finalised models, as shown in Table 4, indicate that no single model can be applied to describe both volume features in either trace. In Keio-I, it is evident that a suitable model for the byte series would be a linear exponential smoothing with 2nd order differentiation ($I=2$) and two MA steps ($MA=2$) with no autoregressive components ($AR=0$), in contrast to the packet series where an ARIMA (1, 2, 3) is more suitable. In addition, Keio-I's byte count modelling differs with the profiling we perform on Keio-II since the Keio-II's byte dataset fits with ARIMA (2, 1, 3).

As opposed to the byte count profiling, packets in both datasets go with the same characterisation scheme under ARIMA (1, 2, 3). The generated outcomes imply that a monolithic blanket scheme

cannot be applied since each raw dataset surely satisfies different assumptions with respect to stationarity, linearity and Gaussianity (Section 4.3). As presented in Section 5, none of the studied datasets (including the WIDE trace) can be sufficiently modelled using an ARIMA process, since all of them exhibit non-stationary characteristics even when differentiated in several orders, thus a non-stationary modelling approach is more appropriate. We also note that model fits estimated by a linearly-based model such as ARIMA depend on whether the initial raw timeseries are linear or not. For instance, the consistency of fitting the exact same ARIMA model on Keio-I/II packets was mainly due to the fact that both datasets have been classified as linear (even though non-stationary) by the Hinich algorithms.

5.5. Protocol-specific characterisation

Despite the fact that volume peaks are successfully detected by the bispectrum, we still lack an interpretation of the main cause of the traffic burst (e.g., which application flows are involved). The main reason for this inability is due to directly dealing with averaged values of the volume aggregate. Consequently, such a scheme poses a level of opaqueness on specifically identifying the flows associated with any particular traffic surge. A more granular methodology which employs protocol-specific modelling allows the identification and extraction of any flows that contribute significantly to the overall burst of the traffic process. Due to the diverse behaviour exhibited by the different protocols as demonstrated by the statistical assumption validation performed earlier (Sections 3.2 and 4.3), we employ the bispectrum on each transport layer protocol and individually assess its dynamics.

5.5.1. Keio-I analysis

In this section we present the analysis performed by the bispectrum only on the Keio-I trace. Fig. 10 graphically represents the bispectral estimates generated for counts of packets and bytes for each protocol, after our series were differentiated in order to remove drifts and noise components. It can be distinctively observed that each protocol exposes certain phase transition peaks on different 2D frequency points, indicating high demands or anomalous behaviour. The bispectral outcomes of the TCP analysis produced consistent characteristics on both packet and byte representations since high peaks due to sudden packet transmissions are normally mapped to high peaks on byte flows. Yet, there are certain phase shifts particularly on the bytes analysis indicating unknown behaviour, which is one extra reason to emphasise the need for separate analysis of byte and packet counts.

Generally in all the three protocols, for both packet and byte-based estimates, there were several distinct peaks within the range of points (0.2–0.4, 0.15–0.4) related to sudden changes caused by a range of applications. The most dominant applications with high frequency mappings were SSH, DNS, HTTP, HTTPS as well as the NNTP protocol. In addition, reasonably high frequency shifts are mainly caused by mail from POP3 and SMTP as well as by RTP streaming. There are particular byte-based estimates having a clear localisation on the time-frequency (TF) plane, but at the same time they cannot be directly linked to the packet frequency characteristics. Especially around point (0.4, 0.2) on TCP, the packet analysis shows a modest peak governing a large area (compared to the rest of the peaks on the same image) whereas the byte-based analysis indicates that the particular region is ruled by phase shifts caused by numerous bytes but not packet transmissions. After mining the raw dataset using the bispectral estimates, we identified an attack targeting TCP port 135 which is used as a service port by the Remote Procedure Call (RPC) protocol. The attacker used various source ports from the same machine to send a large number of single-packet flows with each packet being maximum-sized. Due

to the nature of the single-packet flows, it was not possible to detect this phase transition in the timeseries of packet counts, yet the bispectrum was sensitive enough to reveal it from the byte count timeseries.

UDP applications exhibit greater phase shifts in packets than they do in bytes. Phase transitions observed are consistent with the analysis presented previously (Section 5.2), since their values are within the 2D frequency points ranging from 0.2–0.4 on both the vertical and horizontal axes. With respect to time, there exist parallel phase transitions mainly caused by the BitTorrent protocol (though originally it had been a TCP-based application) that uses UDP for overlay specific actions (e.g. queries between overlay nodes). Phase transitions were also caused by syslog operations used for sending notification messages across the network(s), whereas lower frequency yet still important shifts were issued by applications dealing with DNS and online gaming (mainly an MMORPG protocol). Although a large number of packets was transmitted from each unidirectional flow (resulting in some cases to sudden signal phase shifts), sudden byte peaks are not observed but in one case. By looking at the bispectrum image of the UDP byte counts, the only distinct peak at the point (0.34, 0.31) is mapped to a sudden byte transmission caused by several openVPN-based interactions (UDP port 1194) in our dataset.

ICMP exhibits characteristics on both packet and byte volume features, which were not extracted from the initial aggregate volume profiling. As the representative bispectra show, there are peaks on the expected point regions (i.e. 0.2–0.4 in both axes) having sporadic, scattered shifts. There are three distinct peaks on the packet-level observations at (0.36, 0.28), (0.36, 0.35) and (0.29, 0.35), and five peaks from a bytes perspective. It is quite interesting that one of the packet phase shifts at (0.29, 0.35) is mapped to ICMP inverse mapping. Specifically, an attacker sent ICMP reply packets from a single source to a range of IP addresses behind the Keio university firewall (which did not block them), resulting in a router to respond with an ICMP Destination Unreachable message. This way, the attacker was able to identify the next hop's IP address and expose (a part of) the network's internal topology.

In addition to the inverse mapping, we also identified certain misbehaviours from the bytes analysis. One out of the five peaks, specifically referenced at point (0.4, 0.29) is the incremental behaviour of ICMP flows consisting of a single packet of a length greater than or equal to the typical Ethernet Maximum Transmission Unit (MTU). These flows are known as pings of death, a very well-known attack triggered by ICMP. Apart from this byte-wise anomaly, the rest of the peaks refer to legitimate traffic which was mainly caused by normal ICMP interactions of internal updates taking place within the Keio network initiated by syslog operations.

5.5.2. WIDE Analysis

Fig. 11, depicts the produced bispectra for each volume feature (i.e. bytes and packets) on each transport layer protocol. As evidenced, there was a number of phase transitions on the TF bispectral representation that are correspondingly mapped as sudden traffic peaks on each transport layer protocol in the WIDE trace. By considering all the peaks from all the examined protocols we identify the majority to appear in the range of points (0.2 – 0.4, 0.25 – 0.35) on the bispectral 2D estimate. The protocol-specific analysis has demonstrated a range of intriguing traffic peaks that aided towards a better interpretation of the traffic-wise behaviour in the WIDE trace.

Similarly with the Keio analysis shown earlier and due to the fact that both traces were captured in the same year (i.e. 2006), the traffic trends and peaks were quite similar. In more detail, the most of the identified peaks were due to the most dominant applications of HTTP/HTTPS, SSH, DNS as well as SMTP and NNTP. By contrast with the Keio trace, the WIDE trace also exhibited some

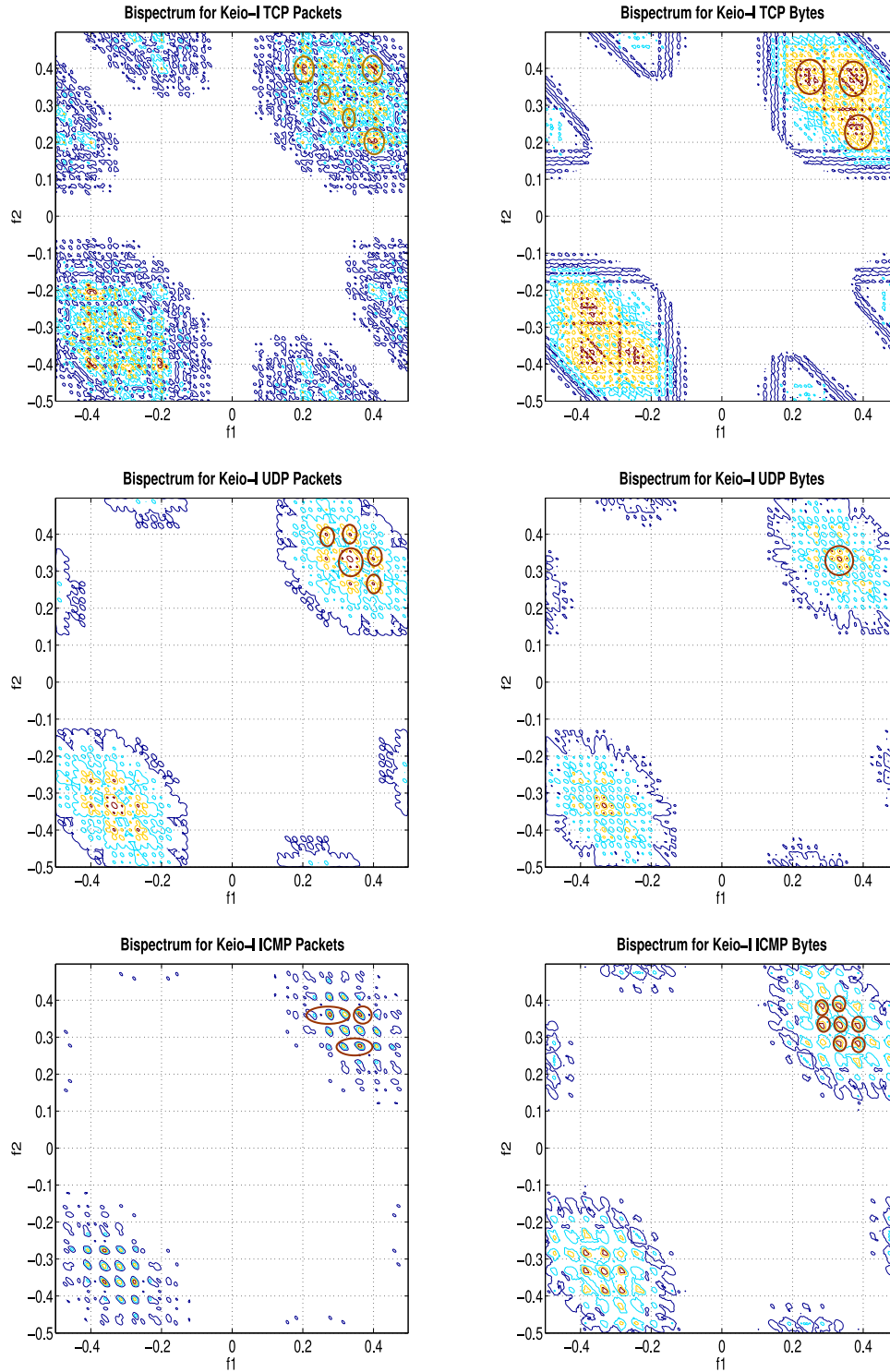


Fig. 10. Bispectra produced after traffic decomposition on Keio-I - packets (left), bytes (right). Significant traffic peaks denoted by red circles. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

distinct peaks caused by the Napster protocol, RTSP and as well as game traffic particularly from games such as Starcraft and Halflife.

TCP applications on a per-packet analysis demonstrated higher amplitudes in their signal phase shifts rather than from a bytes viewpoint. There were three distinct peaks at (0.25, 0.25), (0.3, 0.33) and (0.35, 0.3) that had similar characteristics with respect to the time period of these flows. In fact, all three unidirectional flows had a relatively small duration in the range of 0.3 – 0.5 ms

but with relatively high number of packets. The first two peaks were associated with HTTPS whereas the third peak was related to a large file transfer over the FTP protocol. Interestingly enough, these peaks were not associated with the TCP byte-based phase shifts identified. As demonstrated in Fig. 11, there were two major phase shifts. From a manual inspection we derived that these shifts were associated with unidirectional flows from the same source IP address and they had a relatively small number of packets (i.e.

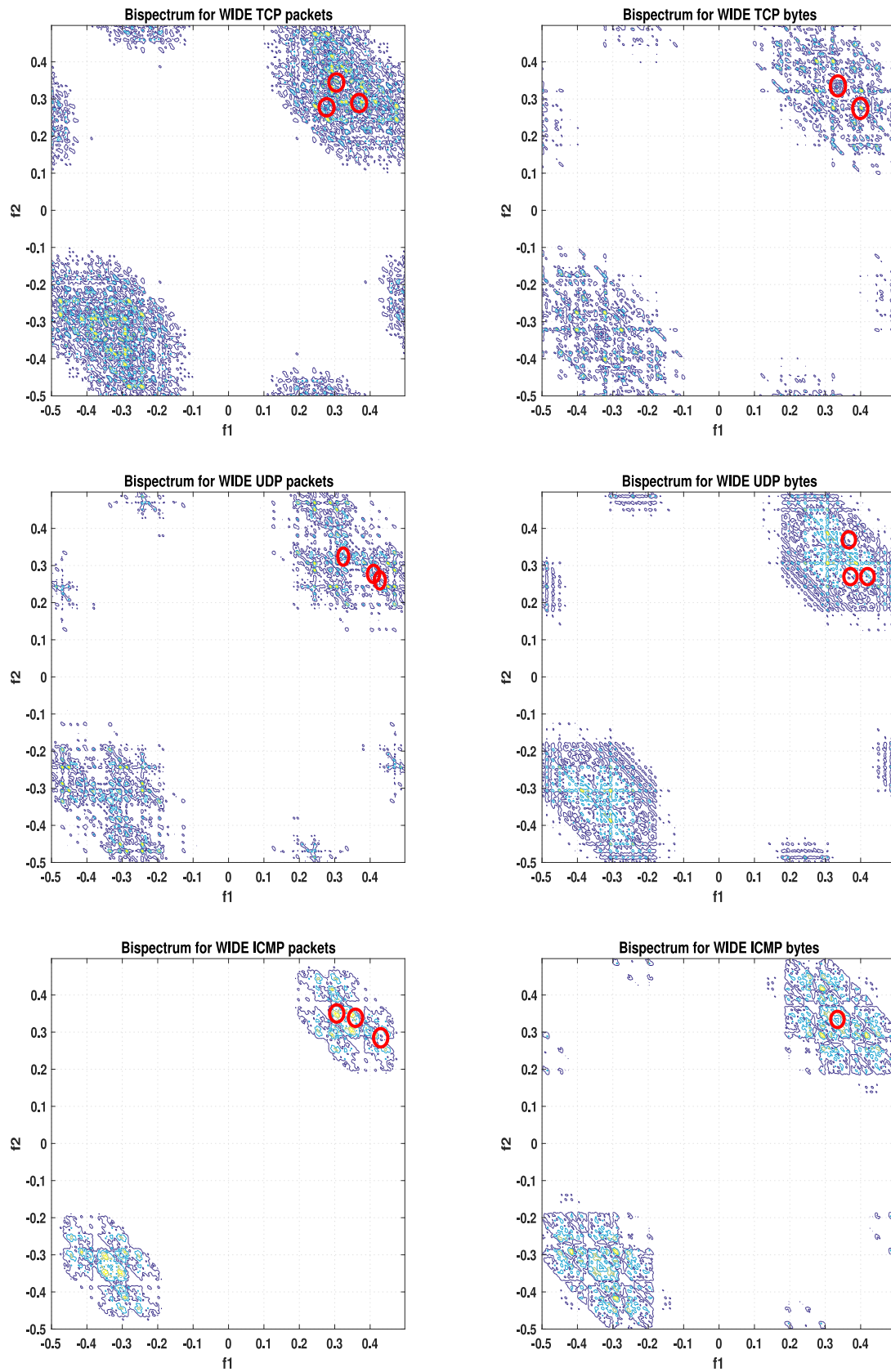


Fig. 11. Bispectra produced after traffic decomposition in WIDE - packets (left), bytes (right). Significant traffic peaks denoted by red circles. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

<20) but with high number of bytes. Most of the packets were reaching the MTU threshold of 1500 bytes/packet and the applications associated were the Squid -an HTTP web proxy protocol- and SSH.

As demonstrated via Fig. 11 there were three significant phase shifts on UDP flows. From the packet view there were shifts in points (0.31, 0.31), (0.4, 0.27) and (0.43, 0.25) whereas in the byte-based analysis were (0.37, 0.26), (0.37, 0.38) and (0.42, 0.26). Two of these shifts on both packets and bytes were related with the same unidirectional UDP flows since they had a quite small duration but rather a large number of packets with relatively high number of bytes for each. A detailed inspection revealed that these flows were triggered by a RealAudio server that was streaming data over the UDP port 6970. The third packet-based shift at (0.43, 0.25) was identified as a single ,packet-wise, large DNS flow in which despite the relatively same size for its packet had an extremely large number of packets with small inter-arrival time. Similarly, the byte-based shift at (0.42, 0.26) was mapped with a large Starcraft flow in which the last 15 packets were byte-wise large and they had a much smaller inter-arrival time compared to the previous packets within the same flow. Thus, a spike with respect to the UDP's byte-wise signal amplitude was localised by the bispectral estimates resulting to the detected phase shift.

From a packets perspective, the bispectra produced for the ICMP protocol revealed three distinct peaks at points (0.3, 0.35), (0.36, 0.33) and (0.44, 0.28). All three peaks were extracted from the raw datasets and they were identified as ICMP scanning activities. The first two peaks were originating from the same source IP address that performed horizontal scans from the port 2048 on multiple DNS domains. Both scans exhibited a large number of ICMP Echo Request packets sent within a small amount of time, thus the amplitude with respect to the TF behaviour of the packet-based timeseries got reasonably high and it was flagged as a phase shift from the resulted bispectrum estimate. The third shift at (0.44, 0.28) was a vertical scan in which a particular source IP address was sending multiple ICMP Echo Request packets on a given destination IP address over multiple ports on the same machine. We argue, that this type of scan had a malicious intent, since the same source IP address appeared to be related with the single peak identified on the byte-based bispectral at point (0.35, 0.34) in which a ping of death was initiated on a different destination host. The two packets contained within that ICMP flow had a size of over 2500 bytes each; hence much bigger than the MTU threshold.

5.5.3. EQUCHAnalysis

The protocol-specific bispectrum analysis on the EQUCHtrace verified that the behaviour of the Internet has changed in terms of which application layer protocols are mostly utilised. In comparison with the analysis conducted on the Keio and WIDE traces that were captured more than a decade ago, the changes related to higher traffic volume and speeds as well as the trends mapping the diversity of application layer protocols dependent on TCP and UDP were quite clear. Nonetheless, as we following describe there were still some similarities, particularly for the analysis performed on the ICMP protocol.

The TCP protocol's volume was the most dominant throughout the trace and, as anticipated, distinct peaks were identified from both a packet and bytes perspective. As illustrated via Fig. 12, there were two distinct peaks pinpointed by the bispectral estimates at (0.32, 0.38) and (0.36, 0.30) that were also related with two out of the three peaks in the byte-based analysis at points (0.30, 0.36) and (0.35, 0.34) respectively. A further post-processing of the raw captured data indicated that these phase shifts were TCP flows in which the extracted IP addresses participated on social network platforms over the usual TCP port 80 via HTTPS and then they

were redirected to YouTube via TCP port 443. Subsequently, the users were assigned to a new connection with YouTube for actually streaming the video content over the TCP port 1935 that is dedicated for the RealTime Media Player protocol (RMTP). This series of events was related to transmitted flows with low duration and in parallel high numbers of packets per flow. In particular, the phase shifts were mostly instantaneous with a large number of users redirected to port 1935 via YouTube. Our analysis could not reveal on whether the redirected users were watching the same video content, however the phenomenon was quite similar with that of a Flash Crowd. Moreover, all the flows associated to these signal phase shifts were peaks that also had a quite considerable count of bytes per packet, hence the redirection of the users to the TCP port 1935 caused sudden increments in the frequency of byte counts. Therefore, a higher amplitude of the byte-based signal on the bispectrum's TF plane was associated with two distinct peaks. The byte-based analysis also shown a third peak at (0.35, 0.36) that also had a quite interesting interpretation in terms of the network's behaviour. In more detail, the third peak was identified to have the behaviour of a typical TCP SYN stealth scan where a single source IP address seemed to perform a mixed vertical and horizontal scan on multiple hosts over multiple ports on a particular destination host. Apparently, the extracted scan didn't hold the typical properties of a normal TCP SYN scan as initiated by tools such as NMAP since the single packet flows had a size of more than 800 bytes each. Therefore, we argue that this scan was carefully crafted and it was probably the probing procedure of a greater botnet.

Of malicious intent were also UDP flows associated with three out of the four phase transition peaks on the packet-based analysis as depicted in Fig. 12. The three packet-based peaks were also linked with one of the two shifts on the byte-based analysis whereas the second byte-based shifts at (0.39, 0.31) was caused by high-byte VoIP flows. In more detail, the packet-based peaks at (0.28, 0.27), (0.31, 0.31) and (0.35, 0.30) were caused by the incremental packet-based behaviour of UDP flows between multiple hosts and a particular source IP address on UDP port 10,050. A further investigation pointed that these flows were in fact initiated through the Zabbix agent protocol that is used for pulling real-time system information and performance metrics from Windows servers and workstations. Due to the real-time nature of the Zabbix agent there were in some cases relatively small duration flows (e.g. <15s duration) with high packet but as well byte rates (e.g. >30K packets per flow and >1300 bytes per packet) that eventually caused the sudden phase shifts detected by the produced bispectra. The fourth packet-based peak at (0.35, 0.35) appeared to be the result of multiple hosts communicating with a single destination IP address through their UDP port 8080. This particular UDP port has been flagged to be used by the Backdoor.Tjserv trojan that acts as a HTTP and SOCKS4/5 proxy and it serves the purpose of listening remote commands from a botserver/botmaster. Thus, we infer that the identified destination IP address could potentially be a command and control server of a greater botnet.

The ICMP analysis demonstrated quite similar characteristics with the other two traces (i.e. Keio and WIDE) in regards of the actual nature of flows that caused the detected peaks. All the three peaks detected for the packet-based bispectrum analysis at Fig. 12 were due to ICMP ping sweeps that are common scanning techniques initiated by tools such as NMAP. In fact, these scans were triggered by a single source IP address over multiple IP addresses over different DNS domains (i.e. horizontal scans) and they were characterized by relatively large packets (i.e. >12 packets per flow) for extremely small flows with respect to their duration (i.e. <5s). The three peaks localized by the bispectrum over the byte-based analysis were all single packet flows with a byte size over the MTU threshold (i.e. >1500 bytes per packet); the so-called pings of death.

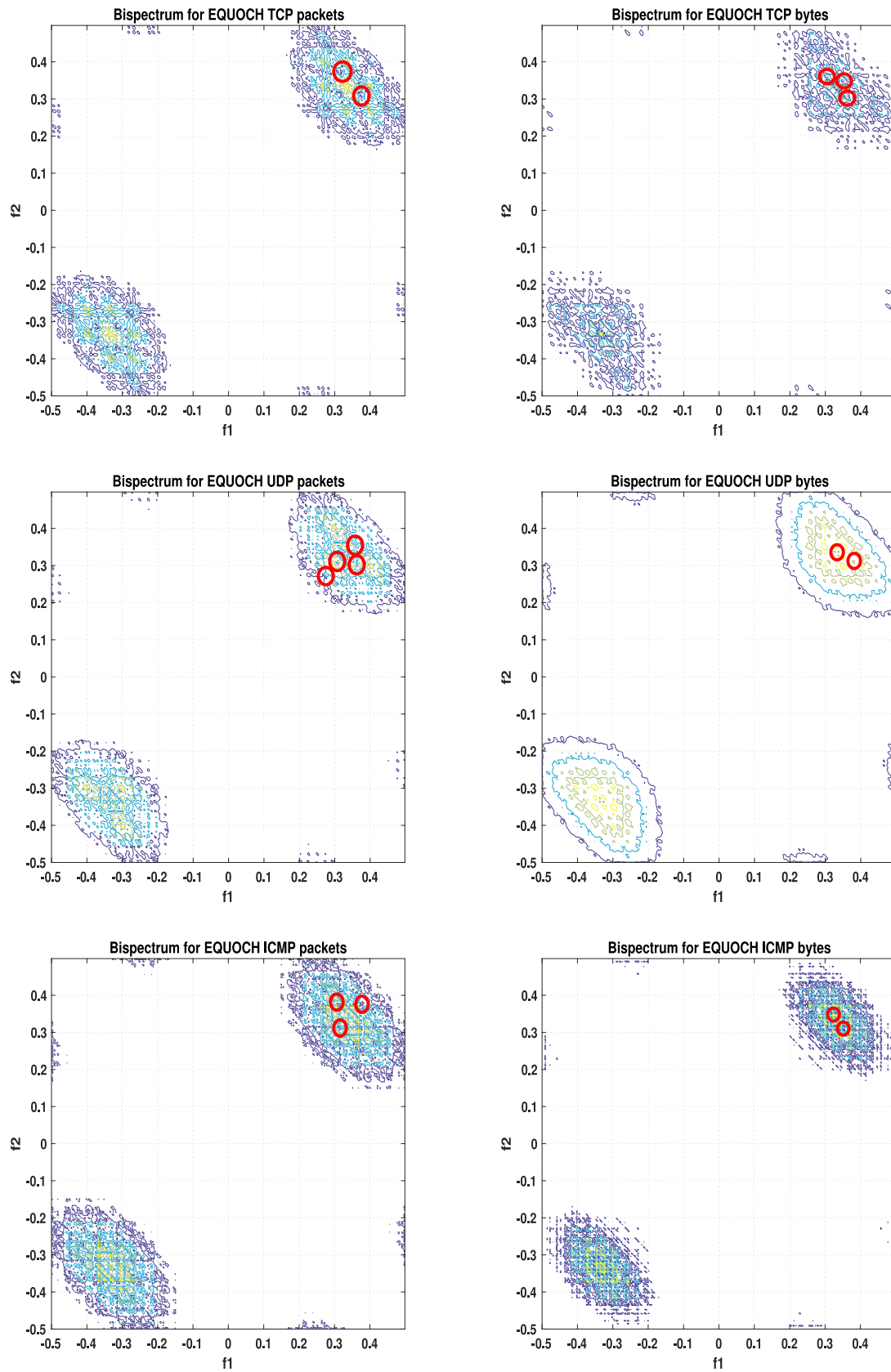


Fig. 12. Bispectra produced after traffic decomposition in EQUOCH- packets (left), bytes (right). Significant traffic peaks denoted by red circles. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

6. Conclusions

Current trends in network traffic profiling include the examination of aggregate volume characteristics, and the use of models that inherently assume *stationarity*, *linearity* and *Gaussianity* of the timeseries. In this paper, we empirically showed that such statistical assumptions do not hold universally, thus they should be rigorously validated beforehand. We proposed the use of Time-Frequency representations for determining stationarity of a time-series, and the Hinich algorithms for validating linearity and Gaussianity. In addition, we proposed the bispectrum for traffic characterisation, and demonstrate its capability of accurately exposing traffic fluctuations by capturing a signal's phase information. The combination of our theoretical and experimental evaluation with our traffic traces collected from backbone and edge links shows that:

- The applicability of any traffic model depends on the underlying statistical assumptions of stationarity, linearity and Gaussianity, which can not be assumed for the entirety of a traffic trace. Hence, such assumptions should be rigorously validated before applying any scheme. Failure to do so results in modelling inaccuracies.
- We propose to use the bispectrum to assess transport-layer protocol or aggregate traffic characteristics, which provides a novel and accurate methodology for revealing application-layer activities and sudden traffic changes. We showed that this capability can be beneficially used within crucial traffic engineering tasks such as capacity planning as well as anomaly and traffic surge detection.
- In contrast to the schemes where only aggregate volume is considered (as e.g. in [2,6,11,13,16,17]), our proposed scheme can also employ protocol-based traffic decomposition to enable a detailed characterisation of the dynamics exhibited by different types of traffic and allow us to extract otherwise hidden patterns such as protocol-specific attacks (e.g. ICMP inverse mapping).
- Link traffic exhibits significant fluctuations in short timescales (30 mins.) which may be described by diverse models (rather than holistic ones, both time and protocol-wise), not adhering to identical statistical assumptions such as *stationarity*, *linearity* and *Gaussianity*.
- Byte counts mostly expose different modelling characteristics with respect to their Gaussian and linear properties than packet counts. Thus, a volume-based approach should consider the analysis of bytes and packets separately.

This work contributes towards the interpretation of traffic dynamics by taking a bottom-up approach and validating the *de facto* assumptions of stationarity, Gaussianity and linearity. It illustrates that these modelling assumptions do change with respect to time and either on an aggregate or protocol-specific viewpoint; thus their validation in the pre-modelling stage is extremely important for the composition of an accurate traffic model. By virtue of the fluctuating traffic behaviour exposed either on aggregate volume or at a protocol-specific level, this paper proposes the use of the bispectrum which can adapt to such dynamics and consequently provide useful insights for crucial traffic engineering tasks.

Acknowledgments

The work has been supported in part by the UK Engineering and Physical Sciences Research Council (EPSRC) projects EP/L026015/1, EP/N033957/1, EP/P004024/1, EP/L005255/1, and by the European Cooperation in Science and Technology (COST) Action CA 15127: RECODIS – Resilient communication services protecting end-user applications from disaster-based failures.

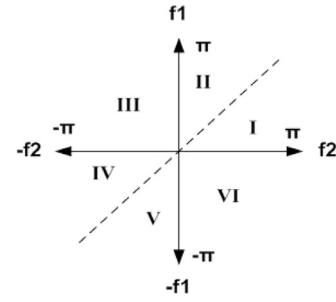


Fig. A1. Bispectrum symmetries.

Appendix A. Bispectrum properties

We explain the symmetrical properties which enable us to correctly relate bispectral estimates with the real events as captured in our raw dataset. Hasselman et al. used the Cramer spectral representations and provided a visualization of the symmetries that exist in the bispectrum, and elaborated upon their importance on real processes [33]. Based on symmetries that exist on the 3rd order cumulant sequence $c_3(\tau_1, \tau_2)$ [34], there is a reflective symmetry for the bispectrum as expressed next:

$$\begin{aligned} B(\omega_1, \omega_2) &= B(\omega_2, \omega_1) \\ &= B(\omega_1, -\omega_1 - \omega_2) = B(\omega_1 - \omega_2, \omega_1) \\ &= B(\omega_2, -\omega_1 - \omega_2) = B(\omega_1 - \omega_2, \omega_2) \end{aligned} \quad (\text{A.1})$$

In addition, for a real process $g(t)$ the bispectrum is equal to its complex conjugate:

$$B(\omega_1, \omega_2) = B^*(-\omega_1, -\omega_2) \quad (\text{A.2})$$

where,

$$-\pi \leq \omega_1 \leq \pi, -\pi \leq \omega_2 \leq \pi, -\pi \leq \omega_1 + \omega_2 \leq \pi \quad (\text{A.3})$$

In simple terms, Eq. (A.2) given the limits for the three independent Fourier components, ω_1 , ω_2 and $(\omega_1 + \omega_2)$ in (A.1) denote that under a real process scenario we may only consider a single region since the rest bispectral estimate domains are symmetric and may be approximated by knowing just one of them. For our case of representing different harmonic Fourier components (i.e. network traffic on different frequencies) in our dataset, we focus on a single region on the bispectrum and identify phase transitions from estimated peaks which are mapped as the exact initiation and time duration of significant volume-wise traffic fluctuations in our datasets. Even though the symmetry relationships provided by Eqs. (A.1) and (A.2) are based on the angular frequency on the TF plane, the symmetry may also be viewed separately in the time and frequency dimension. Fig A.13 illustrates the symmetries expressed by Eqs. (A.1) and (A.2). In our experimentation, we can visualize the bispectral estimates only by observing the values obtained in the positive-valued regions I and II of Fig A.13.

References

- [1] A.K. Marnerides, A. Schaeffer-Filho, A. Mauthe, Traffic anomaly diagnosis in internet backbone networks: a survey, *Comput. Netw.* 73 (14 November) (2014) 224–243. ISSN 1389-1286
- [2] Z. Bozakov, A. Rizk, D. Bhat, M. Zink, Measurement-based flow characterisation in centrally controlled networks, in: *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, San Francisco, CA, 2016, pp. 1–9.
- [3] A.K. Marnerides, A.U. Mauthe, Analysis and characterisation of botnet scan traffic, in: *2016 International Conference on Computing, Networking and Communications (ICNC)*, Kauai, HI, 2016, pp. 1–7.
- [4] D. Brauckhoff, X. Dimitropoulos, A. Wagner, K. Salamati, Anomaly extraction in backbone networks using association rules, *IEEE/ACM Trans. Netw.* 20 (6) (2012).
- [5] H. Kim, et al., Internet traffic classification demystified: myths, caveats, and the best practices, *ACM CoNEXT*, 2008.

- [6] K. Papagiannaki, et al., Long-term forecasting of internet back-bone traffic: Observations and initial models, IEEE INFOCOM, 2003.
- [7] A. Soule, et al., Traffic matrices: Balancing measurements, Inference and modelling, ACM SIGMETRICS, 2005.
- [8] M. Iliofotou, et al., Graph-based p2p traffic classification at the internet backbone, IEEE Global Internet, 2009.
- [9] A. Lakhina, et al., Mining anomalies using traffic feature distributions, ACM Sigcomm CCR 35 (4) (2005) 217–228.
- [10] F. Silveira, et al., Astute: detecting a different class of traffic anomalies, ACM SIGCOMM, 2010.
- [11] M.V.O. de Assis, L.F. Carvalho, J.J.P.C. Rodrigues, M.L. Proença, Holt-winters statistical forecasting and ACO metaheuristic for traffic characterisation, in: 2013 IEEE International Conference on Communications (ICC), Budapest, 2013, pp. 2524–2528.
- [12] Y. Zhang, et al., Fast Accurate Computation of Large-Scale IP Traffic Matrices from Link Loads, ACM Sigmetrics, 2003.
- [13] A. Medina, et al., Traffic Matrix Estimation: Existing Techniques and New Directions, ACM SIGCOMM, PA, USA, 2002.
- [14] N. Groschwitz, et al., G.C., A Time Series Model of Long-Term NSFNET Backbone Traffic, IEEE ICC, 1994.
- [15] J. Cao, et al., Time-varying network tomography, J. Am. Stat. Assoc. (2000).
- [16] O. Goldschmidt, ISP Backbone Traffic Inference Methods to Support Traffic Engineering, ISMA Workshop, USA, 2000.
- [17] A. Soule, et al., Combining Filtering and Statistical Methods for Anomaly Detection, ACM Sigcomm IMC, USA, 2005.
- [18] G. Liang, et al., Maximum Entropy Models: Convergence Rates and Applications in Dynamic System Monitoring, IEEE ISIT, 2004.
- [19] A. Feldmann, et al., The changing nature of network traffic: Scaling phenomena, ACM Sigcomm CCR, Vol. 28, 1997.
- [20] A.K. Marnerides, D.P. Pezaros, H.c. Kim, D. Hutchison, Internet traffic classification using energy time-frequency distributions, in: 2013 IEEE International Conference on Communications (ICC), Budapest, 2013, pp. 2513–2518.
- [21] G. Dewaele, et al., Extracting Hidden Anomalies using Sketch and Non Gaussian Multiresolution Statistical Detection Procedures, ACM Sigcomm LSAD Workshop, Japan, 2007.
- [22] K. Jorma, N. Ilkka, Testing the Gaussian approximation of aggregate traffic, in: Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurement (IMW '02), 2002.
- [23] Y. Zhang, et al., Network Anomography, ACM Sigcomm IMC, CA, USA, 2005.
- [24] C.L. Nikias, et al., Signal processing with higher-order spectra, IEEE Signal Process. Mag. 10 (3) (1993).
- [25] C.L. Nikias, et al., Higher Order Spectral Analysis: A Non-linear Signal Processing Framework, NJ, Prentice Hall, 1993.
- [26] A.K. Marnerides, On characterisation and Decomposition of Internet Traffic Dynamics, 2011 Ph.d. thesis.
- [27] M.J. Hinich, Testing for gaussianity and linearity of a stationary time series, J. Time Series Analysis 3 (1982).
- [28] A. Swami, et al., Higher Order Spectral Analysis Toolbox User Guide, MathWorks Inc., 1998.
- [29] Y. Birkelund, A. Hanssen, Improved bispectrum based tests for gaussianity and linearity, Signal Process. 89 (12) (2009) 2537–2546.
- [30] K. Hasselman, et al., Bispectra of ocean waves, in: M. Rosenblatt (Ed.), Time Series Analysis, 1963.
- [31] H. Akaike, A new look at the statistical model identification, IEEE Transactions on Automatic Control, vol. 19, 1974.
- [32] P. Brockwell, et al., Introduction to Time Series and Forecasting, Springer, 1996.
- [33] K. Hasselman, et al., Bispectra of ocean waves, in: M. Rosenblatt (Ed.), Time Series Analysis.
- [34] T. Matsuoka, et al., Phase estimation using the bispectrum, Proc. of IEEE 72 (10) (1984).



Angelos K. Marnerides is a Lecturer (tenured Assistant Professor) in Computer Networking in the School of Computing & Communications (SCC) at Lancaster University, UK. Dr. Marnerides is an academic member of the Networking Group in SCC, a member of the Security Lancaster Institute and also a member of the prestigious Academic Centre of Excellence in Cyber Security Research at Lancaster University (ACE-CSR). Prior to his current post, he was a Lecturer (Assistant Professor) in the department of Computer Science at Liverpool John Moores University, UK. In the past, he held postdoctoral and senior research associate appointments in the ECE department at Carnegie Mellon University, USA, in DCC at the University of Porto, Portugal and in SCC at Lancaster University, UK. He also held visiting researcher appointments in the SCC, Lancaster University, UK, and the EE department at University College London (UCL), UK. Dr. Marnerides has published widely and has extensive research experience in the area of network resilience with particular interests in network security and smart-grid security for a range of scenarios in the context of the backbone Internet traffic, malware/botnet detection, cloud computing, the Internet of Things (IoT), Advanced Metering Infrastructures (AMI), and vehicular-to-grid networks. His research philosophy is in designing algorithms and systems with strong theoretical foundations and in providing practical data-driven implementations that are deployable in the real world. Throughout his career, his research has been funded by several bodies in the UK (EPSRC, Innovate UK, GCHQ), the European Commission (EC), the US (NSF) and Portugal (FCT). He holds an MSc (Distinction) and a PhD in Computer Science from Lancaster University, UK. He is a member of the IEEE since 2007 and served in the TPC as a member and as a workshop/track co-chair for several top IEEE conferences including IEEE ICC and IEEE GLOBECOM.



Dimitrios P. Pezaros is a Senior Lecturer (Associate Professor) in the School of Computing Science at the University of Glasgow. He is a member of the Glasgow Systems Section (GLASS) and founding director of the Networked Systems Research Laboratory (Netllab) at Glasgow. Dr. Pezaros has published widely and is leading research in computer communications, network and service management, resilience and accountability of future virtualised networked infrastructures, exploring technologies such as, e.g., Software-Defined Networking (SDN) and Network Function Virtualization (NFV). He has received funding in excess of £2.5m for his research by the Engineering and Physical Sciences Research Council (EPSRC), the European Commission (EC), the London Mathematical Society (LMS), the US Federal Aviation Administration (FAA), the University of Glasgow, and the industry (incl. Airbus Group, Brocade, Google, NATS, Solarflare, etc.). Prior to joining Glasgow, Dr. Pezaros has held postdoctoral and senior research associate positions at Lancaster University, where he has worked on a number of EPSRC and EU-funded projects in the areas of network measurement and management, traffic engineering, autonomic communications, and network resilience. He holds a BSc (Hons.) and a PhD in Computer Science from Lancaster University, and has been a doctoral fellow of Agilent Technologies Inc. between 2000 and 2004. He is a chartered engineer, a fellow of the HEA, and a senior member of the IEEE.



David Hutchison is a Distinguished Professor of Computing at Lancaster University and founding Director of InfoLab21. He has served on the TPC of top conferences such as ACM SIGCOMM, IEEE Infocom, and served on editorial boards of Springer's Lecture Notes in Computer Science, Computer Networks Journal and IEEE TNSM, as well being editor of the Wiley book series in Computer Networks and Distributed Systems. He has helped build a strong research group in computer networks, which is well known internationally for contributions in a range of areas including Quality of Service architecture and mechanisms, multimedia caching and filtering, multicast engineering, active and programmable networking, content distribution networks, mobile IPv6 systems and applications, communications infrastructures for Grid based systems, testbed activities, and Internet Science. He now focuses largely on resilient and secure networking, with interests in Future Internet and also the protection of critical infrastructures including industrial control systems.